DOCUMENT RESUME

ED 315 421 TM 014 383

AUTHOR	Black, Harry; And Others
TITLE	The Quality of Assessments: Case-Studies in the National Certificate. Practitioner MiniPaper 9.
INSTITUTION	Scottish Council for Research in Education.
REPORT NO	ISBN-0-947833-34-X; SCRE-107
PUB DATE	89
NOTE	100p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Case Studies; *Colleges; Communication Skills; Educational Assessment; Electronics; Evaluation Methods; *Foreign Countries; Higher Education; Mathematics Tests; *National Programs; *Program Evaluation; Quality Control; Standardized Tests; *Student Certification; Supplies; *Vocational Education
IDENTIFIERS	*Scotland

ABSTRACT

The Scottish Vocational Education Council's National Certificate has its origins in the Scottish Education Department's (SED's) decision to rationalize the provision of certification for non-advanced further education in Scotland under one body. This report presents case studies of assessments of students performed by colleges in Scotland, based on the innovative assessment design associated with the National Certificate. This paper is the second of two reports that discuss the findings of the SED's research project on Assessment in the National Certificate Development Program. Focus is on the quality of assessment for summative purposes; that is, the procedures and instruments that were used to decide whether a student should be credited with having successfully completed a National Certificate module. Topics addressed include the assessment model itself as well as its application to mathematics (in a case study of one college), stock control (in a case study of two colleges), communication (in a case study of one college), and electronics (in a case study of two colleges). The components of the system that seem most responsible for the quality of assessments include the institutions and procedures that have responsibility for the policy at the national level; the process that determines the form and function of module descriptors; and the policies and practices within colleges, departments, classrooms, and workplaces where the assessments actually take place. Twenty-three figures and tables are included. (TJH)

Practitioner MiniPaper

2 Quality and its Determinants

The Quality of Assessments

Case-studies in the
National Certificate

Harry Black, John Hall, Sue Martin and John Yates

The Scottish Council for Research in Education

SCRE PUBLICATION 107

Practitioner MiniPaper 9

ISBN 0 947833 34 X

Printed and bound in Great Britain for The Scottish Council for Research in Education by Russell Print, 14 Forrest Street, Blantyre, Glasgow G72 0JP.

Contents

Figures and Tables

Acknowledgements

None of the work reported here would have been possible without the help and co-operation given to us by the staff of the college departments involved in our research. We owe them a great debt. They freely gave us their time, their efforts and their ideas, despite the considerable pressures that they already face in their work. Our thanks must also go to their students who provided much of the assessment data on which this report is based.

Outwith the case study colleges many other staff throughout the country contributed ideas on assessment instruments, helped us in the process of instrument review and completed questionnaires for us. To all of these we are grateful.

Our Advisory Committee, chaired first by Melville Hendry and latterly by Douglas Law, was a source of much help and advice. SCOTVEC staff provided useful information and guidance.

Our colleagues within SCRE were supportive of the research. Particular thanks must go to Kay Young and Marie Thomas for their clerical efficiency, Dr Sally Brown, John Herdman and Anne Bankowska for their editorial help and the staff of the Research Support Unit for their statistical expertise.

Finally we must acknowledge the help given to us by our funding body, the Scottish Education Department.

Colleges Involved in the Study

Central College of Commerce, Glasgow
Elmwood Agricultural College
Falkirk College of Technology
Glenrothes and Buckhaven College
Kirkcaldy College of Technology
Lauder College
Stevenson College
Telford College

Members of the Advisory Committee

Mr Douglas Law, Kirkcaldy College of Technology (*Chair from January 1987*)
Mr Melville Hendry, Scottish Examination Board (*Chair until January 1987*)
Mrs F Hope Johnston, Scottish Education Department
Mr David Kelso HMI, Scottish Education Department
Mr Jim Kennedy, SCOTVEC
Mr Douglas Thomson, Dean Education Centre
Mr Bert Whiteside, SCOTVEC
Dr Sally Brown, SCRE

Preface

It is now five years since the National Certificate was first implemented in colleges and schools in Scotland. The changes in assessment procedures which it brought in its wake have had many implications for teaching staff throughout post-compulsory education. The research project on *Assessment in the National Certificate Development Programme* was set up to examine the implications of these changes.

This report complements our earlier publication *Assessing Modules* (Black, Hall and Yates, 1988) which examined the views of staff teaching National Certificate modules in a group of colleges throughout Scotland. The present volume is concerned with the technical qualities of the assessments which are actually made by staff.

We set out to ask about the quality of the assessments which were taking place in colleges and to identify those factors which might affect that quality. Such an investigation must, of necessity, deal with some apparently esoteric areas of assessment technology and, to that extent, some parts of the report may appear daunting to those unversed in such matters. However, we have tried to confine these technicalities to Chapter 2, where we discuss the assessment model which informed our research. We hope that the chapters on our case studies will be more accessible to the reader who has no taste for statistics and that they may find in them something which they can apply to their own practice.

There are two important features of this research which should be borne in mind by the reader. The first concerns the implications of the research design we used. Because we opted to look in depth at a few modules in a limited number of settings we cannot claim that the findings we present are generalisable across the country. What we would say is that the modules studied were chosen to represent key competences and were likely, in our view, to reveal the types of problem common to certain forms of assessment. We hope that the result is an illuminative insight into the factors which affect the quality of assessment within the National Certificate.

The second point we would wish to make is that some of this research has already passed into history. We looked at assessment in modules delivered in sessions 1986/87 and 1987/88 but development of the National Certificate continues apace. We are aware that SCOTVEC, for instance, has a rolling programme of module review and development and that the system is continually being refined, at both national and local level. Nevertheless, we feel that many of the points we make have a general relevance which extends beyond the details of the particular modules examined.

As we say at one point in our final chapter, assessment technology has its limitations and human beings are fallible. This being so we cannot offer solutions to all the problems posed by assessment. We would hope, however, that this report would go some way towards clarifying some of those problems and that staff in colleges, schools and elsewhere will find the insights it offers helpful in reflecting on their own practice.

Appendices relating to the interview schedule, questionnaires and associated results, together with the module descriptors for the four case studies have been compiled as a separate document, available from SCRE.

1 Background

The assessment system used for the Scottish Vocational Education Council's (SCOTVEC) National Certificate is both radical and topical. It is radical in its use of a criterion-referenced model, in the responsibilities it gives to teachers and in the demands it makes for quality control. It is topical because it is in the vanguard of a number of assessment systems for modular curricula which are a feature of many current educational developments. These include much of the innovative work associated with the Technical and Vocational Education Initiative (TVEI) throughout the UK and the work of the National Council for Vocational Qualifications (NCVQ) in England and Wales. Because it places new demands on staff and because of its central importance in certification of post-compulsory education, the Scottish Education Department (SED) commissioned a research project on Assessment in the National Certificate Development Programme. This is the second of the two reports which discuss the findings of that project.

THE NATIONAL CERTIFICATE

The National Certificate had its origins in the SED's decision to rationalise the provision of certification for non-advanced further education in Scotland under one body. Until the middle of the 1980s, certification of courses in this sector had been provided by the Scottish Business Education Council, the Scottish Technical Education Council and other UK agencies such as the City and Guilds of London Institute, Pitmans and the Royal Society of Arts. However, it was believed that this proliferation of certificates resulted in confusion amongst students and employers as well as posing administrative difficulties in colleges.

Accordingly, in 1983 the SED published *16-18s in Scotland: an Action Plan* (SED, 1983) which proposed the establishment of a single body to be responsible for the accreditation of all such courses. The new provision was to be in modular form, and by the time it was implemented in session 1984/85 there were some 600 modules available. By March 1985, when SCOTVEC was established, there were 1700 modules in the catalogue, and this has now grown to around 2500. At the time of writing, these modules form the basis of course provision throughout the non-advanced further education sector in Scottish colleges, and they are also used in many schools as part of the curriculum for both 14-16 and 16-18 year olds (SED, 1988).

THE ASSESSMENT MODEL

Our first report on this project (Black, Hall *and* Yates, 1988) identified a number of key features of the National Certificate assessment model. These included focus on a description of attainments rather than a measure of general ability; the prescription of goals in the form of pre-defined performance criteria; and its almost exclusive reliance on college- or school-based assessment during the teaching of the module. It was also noted in the report that some local authorities and colleges had felt the need for local quality control systems to supplement the national system provided by SCOTVEC. In brief, therefore, the National Certificate is based on a criterion-referenced internally-assessed model with quality control exercised centrally by SCOTVEC but supported in some cases by local systems.

A NATIONAL CERTIFICATE MODULE

The modules available are listed in a 'catalogue' which is divided into nine sections covering Interdisciplinary Studies; Business and Administration; Distribution Studies, Food Services and Personal Services; Engineering; Built Environment; Caring; Industrial Processing; Land and Sea Based Industries; and Pure and Applied Sciences. However, the catalogue itself lists only the module titles. To understand how assessment is built into each module it is necessary to turn to the 'module descriptor'.

Each module in the catalogue has a descriptor which may vary in length from about four pages to, in some instances, including appendices, more than 60. However, irrespective of its length, each descriptor is built around the common sections outlined in Figure 1.1.

For this study, the important sections are 5, 6 and 8. Under Learning Outcomes, section 5, the descriptor offers a statement of

Figure 1.1 The Structure of a Module Descriptor

Module Descriptors have been designed as Curricular Frameworks consisting of nine sections:

1. REFERENCE NUMBER and DATE
2. TITLE
   to give a clear idea of what the module is about
3. TYPE AND PURPOSE
   to give a clear, detailed guide to the uses for which the module was designed, the ways in which it can best be used and any limitations on its use or recognition.
4. PREFERRED ENTRY LEVEL
   to show the level of previous achievement without which it is likely that a student will have difficulty in successfully completing the module.
5. LEARNING OUTCOMES*
   to specify unambiguously the key competencies resulting from the successful completion of a module.These cannot be changed.
6. CONTENT/CONTEXT
   to give an indication to tutors of the subject matter which would assist in the achievement of the Learning Outcomes.
7. LEARNING AND TEACHING APPROACHES
   to suggest learning strategies which enable the Learning Outcomes to be achieved in as student-centred, participative and practical a way as possible.
8. ASSESSMENT PROCEDURES
   to show in detail what the student must do, and to what level, in order to show that the Learning Outcomes have been mastered. Recommended Assessment Procedures may not be altered without the prior approval of the Council.
9. EXEMPLARS AND GUIDELINES
   these are sometimes included to give tutors additional support by way of background information and examples of assessment material.

Source: The National Certificate Catalogue of Module Descriptors. SCOTVEC 1987-88

**in current module descriptors this section also lists the performance criteria*

the behaviour, skills and knowledge on which the student will be assessed. These are written in behavioural terms but are not always sufficient in themselves to constitute a domain definition. The term 'domain definition' can be thought of as the statement clarifying what can be legitimately included in the assessment instrument, and indicating to *users* of assessments (students, employers) what has to be mastered. For a discussion see Black *and* Dockrell (1984). Section 6, Content/Context, offers further clarification by indicating legitimate content of the domain, while Section 8 identifies what a student must do to be considered competent. In all cases 'mastery' is construed in dichotomous terms: the student 'has' or 'has not' mastered the outcome, but there is considerable variation amongst modules as to how this is defined.

The National Certificate model has adopted much of the 'state of the art' in criterion-referenced assessment design (Popham, 1978; Berk, 1980; Roid and Haladyna, 1982). However, establishing a sound model on which to build assessments does not necessarily lead to sound assessments. Furthermore, in moving bevond relatively familiar areas such as Maths and Technology to less researched areas such as Personal and Social Development, the National Certificate is breaking new ground in criterion-referenced assessment.

RESEARCH QUESTIONS

The innovative assessment model of the National Certificate determined the questions for the research. In our earlier report we focused on the views of teaching staff and assessment for its broad range of purposes. In this report we concentrate on assessment for summative purposes - that is, the procedures and instruments which the staff used to decide whether a student should be credited with the module.

The two basic questions we asked were: *how sound* were the assessments, and *what might explain their quality?* What do we mean by 'sound' and 'quality'?

By 'sound' we mean that the assessment is fair and defensible. For the National Certificate it must provide an adequate statement about the specific learning outcome it purports to measure. It should allow the user to arrive at clear and accurate decisions about students' attainments. It should be a reliable indicator to end-users (employers, staff in other educational establishments, or students) of the particular message the National Certificate aims to deliver. And it should be carried out in a professional and unbiased way, which supports sound practice, not only in assessment, but also in pedagogy. Our understanding of the basis on which quality, or the lack of it, might be explained also has its origins in our own perspectives but it is supported by more systematic analysis.

One of the first tasks for the team was to answer a supplementary question. This required us to review the literature on assessment, particularly criterion-referenced assessment and the documentation on the Action Plan, and to decide what factors might influence the adequacy of assessments in the National Certificate. The outcome was a model for analysing the assessment system which is described in some detail in Chapter 2. The two basic questions about how sound assessments are and the reasons for their quality

are interpreted in a number of different ways throughout the rest of the report to meet the needs of the specific studies.

METHODOLOGY

From the National Certificate literature, the general literature on criterion-referenced assessment and the findings from the first phase of our work on staff perceptions (Black, Hall *and* Yates, 1988) we identified factors which might influence the quality of assessment. These formed the basis of the 'quality of assessment' Q-model applied first to our case-study in Mathematics and then to three other case-studies. At the same time, we carried out a national survey of staff views on these and other modules. Finally, the findings from the four case-studies and the national survey were synthesised into an overall comment on the quality of the assessments we encountered and a reflection on the policy implications this might have.

It is important to consider the nature of our case-studies and how our findings might be interpreted. Time and resources were finite and National Certificate modules are taught in a very large number of centres. We could look at assessment in the detail we felt appropriate in only a small number of colleges and so adopted a case-study approach. It follows that while we know a lot about the quality of assessments in the four cases on which we worked, we cannot claim that what we found is representative of National Certificate assessment as a whole. What we can say is that there is a range of quality at least as great as that we encountered, and that, in relation to the criteria we had established, there was substantial variation among our case- studies. The detailed knowledge we have of the reasons for this variation in quality allows us to generate some hypotheses of more general application but these should not be confused with the findings which might have been generated from a larger scale study.

We chose case-studies which were very different in nature: modules from Mathematics, Stock Control, Communication, and Electronics. Within this group there are 'practical' and 'academic' subjects and a range of assessment instruments and module descriptors, some of which we considered provided precise domain definitions while others are less so.

Because of this variation, each study required a different approach but there are common threads linking them. Staff involved in teaching the modules in the case-study colleges were interviewed (Appendix 1)* and each completed a questionnaire (Appendix 2). The purpose of these was to explore their perceptions of the module and aspects of their practice to help us understand the reasons for the quality of the assessments being made.

Each study also involved the collection and detailed analysis of data about the assessments which staff were making as part of their normal practice in their classrooms or workshops. In all but the Electronics study, there was an element of comparison of instruments or procedures within one college, between colleges or between elements of both of these.

---

*The Appendices mentioned in this report are contained in a separate document ('Quality of Assessments - Appendices') available from SCRE. 'Annex' 1 is included at the end of this document.

Finally, the findings from each study are supported by data from a questionnaire to staff teaching the module elsewhere in Scotland (Appendix 3).

THE REPORT

Chapter 2 describes the criteria used to comment on the assessments and the factors we expected to influence the adequacy of assessments. Together these allowed us to build the 'quality of assessment' model which is described in the final section of the chapter. This chapter provides the necessary theoretical background to the research.

Chapters 3 to 6 offer separate accounts of the four case-studies. The Mathematics study (Chapter 3) was primarily concerned with using alternative assessment instruments, devised outwith the case-study college, and examining the assessment decisions for comparability with the college-devised instruments. In the Stock Control study, reported in Chapter 4, we made use of alternative assessment instruments devised by a case-study college and also exchanged college-devised instruments between colleges, again with a view to studying the comparability of the assessment decisions.

Both the Mathematics and Stock Control studies deal with fairly formal types of testing, using short answer and multiple-choice questions. In Chapter 5 we look at assessment within the Communication modules, which is less formalised, and where the skills assessed are more difficult to define accurately in advance. Short sub-studies examine also the assessment of Writing, Talking and Listening.

The Electronics study reported in Chapter 6 is different from the others. Here we were primarily concerned with the assessment of process skills in a practical workshop situation. To gain insight into how such assessment would work, one of the research team joined a class for the duration of the module. Chapter 6 presents his observations and reflections on this experience.

Finally, in Chapter 7, we draw our findings together and consider what our case-studies reveal about the quality of assessment within the National Certificate. We offer comments on the implications this might have for policy-makers and indicate some questions which are worthy of further research.

2 Quality and its Determinants

Assessment should be judged in relation to a particular set of requirements. In some cases, for example, assessment would be considered to be of high quality if it sorted those tested into a normal distribution; in others, quality is judged by how well an assessment predicts success at some later stage; in some applications of diagnostic assessment it will be of high quality only if it pinpoints accurately the reasons why a student is having learning difficulties. What then are the requirements against which National Certificate assessments should be judged?

The best starting point is the description of the Action Plan and its subsequent application by SCOTVEC. Three principal aims of assessment are given (SED, 1983; SCOTVEC, nd):

- to ensure national standards
- to indicate to students successful learning and areas requiring further work
- to provide feedback to teachers on teaching and learning approaches and individuals' problems

These functions are fulfilled by formative and summative assessment. Assessment in the National Certificate Development Project has concentrated on summative assessment and thus on the first two functions. Summative assessment is said to be the 'basis upon which it is decided whether the student has attained the necessary national standard ... and therefore has to be a fair and valid judgment of the student's performance'. (SCOTVEC, nd p1).

Further details are given on two essential components of summative assessment - validity and concordance with a national standard. The assessment should 'match as closely as possible what has been learned with the statements in the module descriptor relating to the learning outcomes, instrument specifications and levels of expected attainment' (SED, 1983). Not only should the assessment be a fair test of the learning outcomes, it should also be consistent with the national standard, regardless of where a student has taken a module, so that 'testing of students in Centre A can be seen to be as relevant and accurate as students following the same module in Centre B' (SCOTVEC, nd p7).

In this report the most crucial criteria on which the quality of assessments in the National Certificate are judged are taken to be the extent to which

1 they are valid descriptions of attainment of the given learning outcome;

2 they produce accurate decisions about whether a student has or has not satisfied the performance criteria appropriate to the learning outcome.

WHAT DETERMINES THE QUALITY OF ASSESSMENTS?

Our ideas about the determinants of quality, derived from the literature on assessment and the first stage of the project, provided a foundation for the first (Mathematics) case-study. Since this seemed to provide a suitable framework within which to proceed,

we developed it into a 'quality of assessment model' (Q-model) and used it to generate questions for the later case-studies.

The general literature on assessment was helpful in identifying such technical characteristics of good instruments as the language used and the design of 'distractors' in multiple-choice testing. However, most of it, particularly that from British sources, is written around the priorities of the norm-referenced systems of accreditation of the last two decades and was of only marginal relevance.

The literature on criterion-referenced assessment was more fruitful for understanding the technical considerations which determine quality. Amongst the factors identified were:

- the adequacy of the domain definition in the module descriptor;
- the clarity and appropriateness of the way in which mastery is conceptualised in the performance criteria for each learning outcome;
- the adequacy of the guidelines provided on how to construct assessment instruments;
- the adequacy of the instruments in actual use for making assessments.

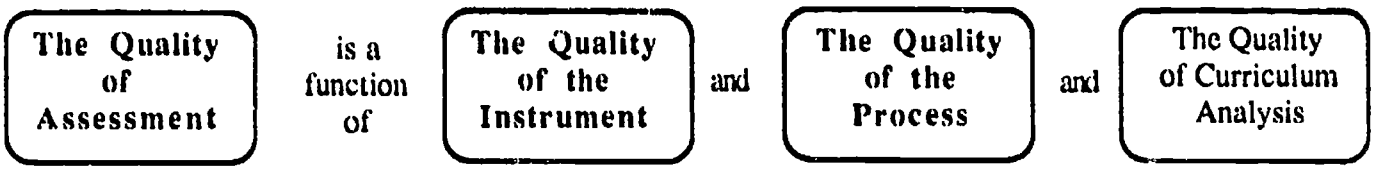
However, another set of factors, to which there was little reference in the literature, had emerged as important from the first stage of our research. These related to the process of assessing, rather than the nature of the instruments used. Their importance is heightened by the 'new' role of the teacher in National Certificate assessment. It was clear that if process factors, such as the following, were not included in any attempt to understand the quality of assessments, only a partial view would be obtained. These are:

- attention to fundamental 'professionalism' in administering assessment;
- attention to performance criteria when making mastery decisions;
- commitment to the assessment philosophy which underwrites the Action Plan.

We also identified factors likely to influence the quality of assessment but outside the control of our work. These included variation in the quality of curriculum analysis which underwrote the module descriptor. If the learning outcomes identified by the module writers were seen by teachers to be relevant and appropriate they would be more likely to take them seriously and assess them well. However, systematic analysis of this factor was beyond our resources. Similarly, we had earlier established that effective quality control at both a local and a national level is essential. However, as we concentrated on in-college case studies, we could gather no systematic evidence on the effectiveness of subject assessors and college moderators.

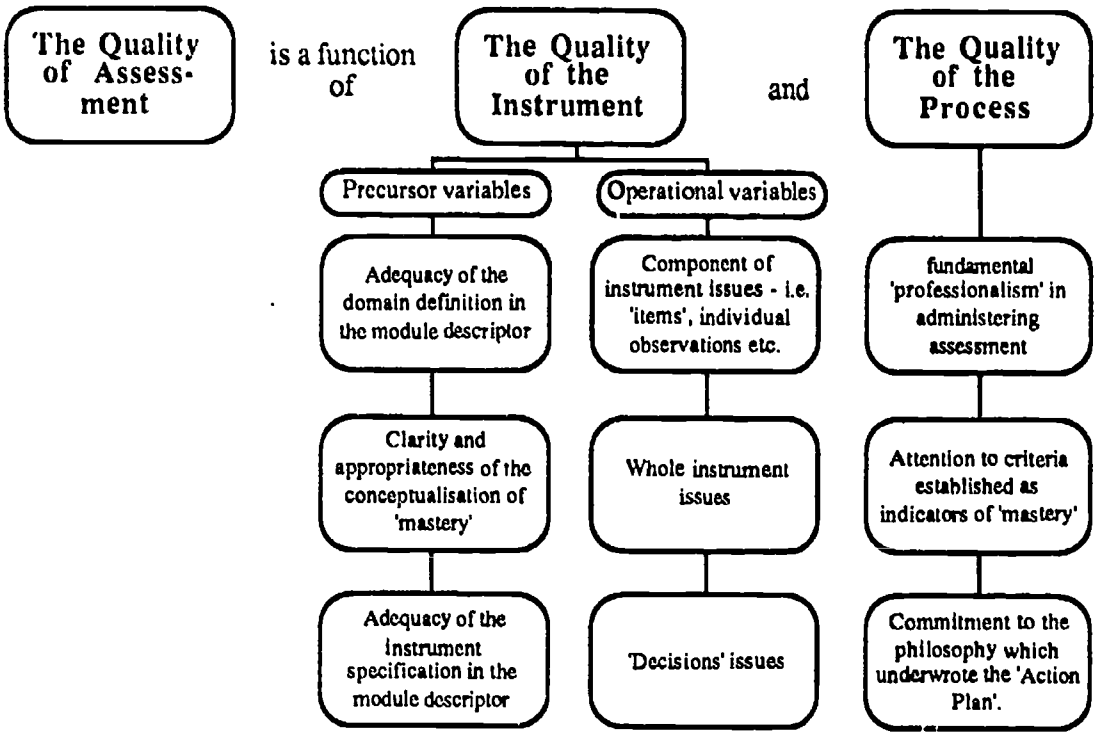
THE Q-MODEL

At its simplest, the Q(uality)-model suggests that, within any given context, the quality of assessment is dependent on the nature of the instrument, the way in which the assessment is carried out and the quality of the curriculum analysis to which it relates.



To be of practical value, however, more elaboration is required, and this is offered in Figure 2.1. The 'Quality of the Instrument' components fall into two categories. One set of 'precursor' variables includes the quality of the module descriptor and any other support given to the teacher prior to constructing his or her assessments. The second set of 'operational' variables relates to the quality of the resulting instruments themselves the mastery decisions arrived at by using them. Three questions are asked in relation to the 'Quality of the Process' component. Their origin lies in the first stage of the project.

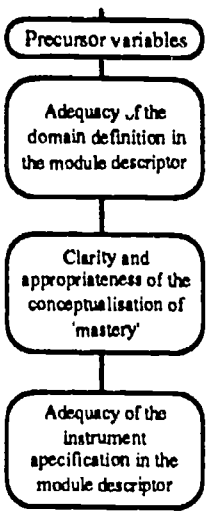
Figure 2.1 The Major Features of the Q-Model



The specific questions which we considered appropriate to ask in relation to each element of the model are set out together in Annex 1. Not all of these questions could be answered in the same way for each case-study.

Instrument Issues: Precursor Variables

Adequacy of the domain definition



The quality of the domain definition is of central importance in determining the quality of criterion-referenced assessments. In National Certificate modules this is manifest in the statement of the learning outcomes and the content/context for assessment, although we found that staff developed their understanding of the domains from the whole descriptor and not only these specific sections. The importance of the performance criteria in helping to define the limits of a domain was particularly evident in some cases.

Our experience suggests that there are three questions which need to be asked about a domain definition. The first is how clear is it what the individual being assessed will have to do, in other

words, what skills or competences are involved. The second is a question about the 'content limits' - is it clear in what range of circumstances the skill or competence is to be applied? And finally, it is important to ask whether the domain is of a realistic size. If the domain is very broad, it can be difficult to design sound instruments to assess it. And even if it is possible to design instruments for broad domains, they are likely to be very lengthy.

Clarity of what is meant by mastery

In previous work on criterion-referenced assessment, there is only limited discussion of the concept of 'mastery'. The reports on scoring criterion-referenced tests concentrate on multiple-choice design, with little about other modes of assessment such as the practical tests, projects, or extended writing which characterise the National Certificate. From the assessment perspective, therefore, the approach may be breaking new ground.

Our questions must reflect this range of assessment approaches. First, is 'mastery' clearly defined in the module descriptor, and is this definition congruent with the behaviours in the domain definition? For example, if a domain involved applying a particular concept in a practical situation, it is unlikely that testing for recall would provide assessments of high quality. The key question is whether the module descriptor offers a clear understanding of what distinguishes students who master a learning outcome, from those who do not. Where mastery is defined by a cut score (eg 80%, 70%) we would also look at how the cut score is justified.

Adequacy of instrument specifications

Although others have made little distinction between 'domain' and 'instrument' specifications, we have found that most domains can be assessed by a range of assessment instruments. Because module descriptors are explicit in offering advice on how to construct instruments, the final 'precursor' component of our model covers this area.

There is substantial variation in the guidance given in the module descriptors on the construction of instruments. In some cases this is accompanied by 'exemplars'. These exemplars could result in technically sound assessments but equally they could tempt staff and students to rely on them as the only mode of assessment. On the other hand, in cases where only minimal support is given to teachers on instrument construction, much is demanded of their understanding of the requirements of assessment. The most appropriate solution might be a module descriptor providing a sound domain definition and a sufficiently flexible instrument specification to result in a variety of sound forms of testing.

The questions included in the Q-model concerning instrument specifications reflect this position. We see the quality of guidance given on appropriate instruments, on the appropriate length of the test and on basic construction techniques, as having a bearing on quality. Overall, however, we must ask whether the specification will support the construction of instruments which yield comparable results.

The effect of the precursor variables: 'tightness' of definition

These precursor variables led us to distinguish between those modules which we classified as 'tightly' defined (on the basis of the information in the module descriptor) and those 'loosely' defined.

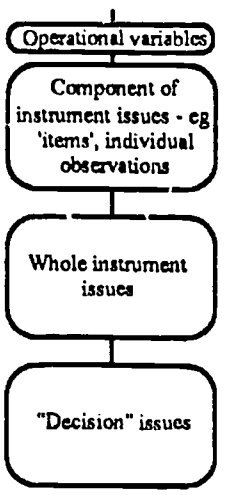
We expected that a module descriptor containing clear and unambiguous statements of the learning outcomes, well-defined content limits (and not too much content), unambiguous statements of the performance criteria to be met, precise specifications for assessment instruments, and appropriate exemplars, would be likely to result in high quality assessments. Knowledge of what was to be assessed would increase the validity of assessments, while knowledge of how decisions were to be made would increase the reliablility of assessment. Such descriptors are 'tightly defined'.

In contrast, a 'loosely' defined descriptor would contain learning outcomes, content limits and performance criteria which were ambiguous and open to various interpretations. The content might cover a large area from which to select for assessment purposes. Guidance on assessment instruments would be only rudimentary and there might be over-reliance on the use of the lecturer's professional judgment. It could be possible for two lecturers, teaching the same module, to assess their students on different knowledge or skills, using different assessment instruments, and basing their assessment decisions on different ideas of what demonstrates mastery. This would not be high quality assessment.

Of course, the nature of the subject matter has an influence on how 'tightly' a module descriptor can be defined. Caring Skills will not be as tightly defined as Typewriting. It is not always easy to provide clear and objective statements to delineate a subject and define what mastery entails. We expected different subjects would have more, or less, 'tightly' defined descriptors and that this would affect the quality of the assessments and the types of problem we would encounter.

Instrument Issues: Operational Variables

There is overlap between questions arising from the precursor and operational variables of the model, but this is not duplication. While precursor variables reflect the information and support available prior to constructing assessment instruments, the operational variables only become available for inspection once the assessment instruments have been produced. The questions are, therefore, about the quality of the product which arises from the guidance given in the module descriptor.



These questions can be about the quality of three aspects of assessments: components of instruments; whole instruments; and decisions. Our model considers each of these in turn.

Component of instrument issues

At the most precise level, it is the quality of the individual items in a test, the individual assessments related to the criteria used in assessing written work, or the individual observations made in the practical setting which determine the quality of the assessment instrument. Quite a lot is known about evaluating individual items

in multiple-choice tests, but very little as one moves away from that mode of assessment. The principle of considering the quality of individual components, however, is appropriate irrespective of the instrument being used.

There are many features of component items which might influence their quality, and we can be explicit about only some of them. The first feature is the relationship between the component and the domain definition. Is the component congruent with the domain definition in context, behaviour and content?

The second characteristic concerns the extent to which the component item is a true measure of the domain, or part of the domain. Almost every form of assessment comprises a measure of what is being assessed mixed with extraneous 'noise'. This can arise from a variety of sources including, for example, the language in which the item is presented and the extent to which it assumes other knowledge and skills. Such noise has to be kept to a minimum.

A third characteristic of components is the extent to which they discriminate between masters and non-masters. In a criterion-referenced context, distinctions are based on a simple dichotomy between masters and non-masters.

Our final question in this section is more general, and asks whether the instruments have been reviewed. This review can include both systematic consideration of quality through inspection, and empirical processes, based on statistical procedures which describe and analyse the results of applying items. Component items which have been the focus of review are more likely to yield assessments of quality than are those which have not.

Whole instrument issues

Whole instrument issues relate to the 'sum of the parts' which go to make up an assessment instrument. Like 'component' issues, these include 'freedom from noise' and whether instrument review has taken place. The procedures may be different in these two contexts, but the principles remain.

There are several other characteristics of whole instruments which have a bearing on the quality of assessment. First, it is important that the domain has been adequately sampled. An instrument may comprise a set of components all of which are adequate but which is itself, an inappropriate measure of the domain. Only if the instrument covers all aspects of the domain can it claim to be a valid measure of it. This has implications for the length of test or the number of observations required to produce a sound assessment. If a domain comprises a single clearly-defined skill, concept or element of knowledge, it is possible to produce comparable assessments from several short tests relating to that domain. However, if the domain is larger, or can be broken down into disparate elements, instruments must be correspondingly longer if comparability is to be maintained and equivalent classifications of students made. The philosophy of the National Certificate is clear that different staff in different colleges should produce instruments yielding comparable assessments of student mastery, and so the length of the assessment instrument must adequately reflect the scale of the domain.

The extent to which a variety of assessment instruments measure what they were intended to measure is indicated by the extent to which they classify 'masters' and 'non-masters' consistently. If the domain is small, good instruments assessing the same outcome will lead to similar decisions on mastery. If they do not, either they are assessing different things, or there are unacceptable levels of noise.

Finally, a variable which exists on the margin between the 'instrument' and 'process' components of the Q-model is the issue of practicability. Even technically sound assessment procedures will only yield quality outcomes if they are practicable in the classroom or in the workplace.

Decisions issues

Assessment instruments yield data on which decisions about mastery can be made, but staff must interpret these data to decide how to allocate students. The procedures they use have a substantial bearing on the quality of the process.

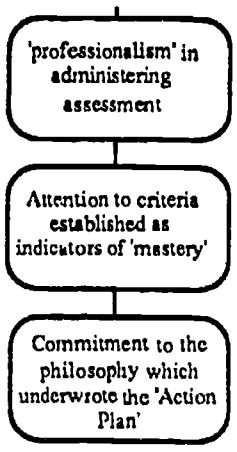
Our first question asks whether the way that a mastery decision is made is congruent with the chosen definition of mastery. If, for example, a cut score for mastery is set at 80%, the reliability of assessments will be reduced if staff vary from this. More subtly, if the assessment is directed at *application* of a skill but mastery decisions are based, for example, on *knowledge* of the skill, the quality of assessment will be placed in jeopardy.

There are also important questions about the stability (reliability over time) and equivalence (reliability between alternative forms of assessment or among assessors) of assessments. A sound assessment should yield similar results if applied to the same students at different times (assuming no further learning has taken place) and similar mastery decisions to any other instrument drawn from the same domain definition. There is seldom a 100% agreement between alternative instruments or repeated measures, but a low level of agreement would indicate that something is wrong.

Our final question is the most difficult to deal with in the National Certificate context. It is well-established that almost any criterion-referenced assessment can only provide an *estimate* of the 'true test' score, because it only samples the domain. There is a substantial technical literature on how the 'observed' score can be translated into a 'true' score. However, statistical techniques to estimate true scores are complex, sometimes rely on assumptions which do not hold good in our context and may require prior knowledge about the students or the tests, which is unlikely to be available. Where domains are reasonably small in scale and assessed by a sufficient number of items or observations, however, the difference between 'true' and 'observed' scores is typically not great.

Process Issues

Many of the process components of the Q-model are grounded in our work in the first phase of the project. They relate principally to questions about how staff carry out assessment but also include questions about how staff have interpreted the philosophy which underwrote the Action Plan.



Professionalism in assessment

There are certain assumptions about how teachers will act in relation to any assessment. At the most extreme, if teachers were to give students the answers while testing was taking place, there would be no value in the resulting 'mastery decisions'. While this extreme is likely to be rare, there can be substantial differences in interpretation of the official assessment procedures.

For example, questions can arise over whether students have been given equal opportunities to display mastery. The instruction in the National Certificate literature to allow students to continue working on a learning outcome until they show evidence of mastery is open to a number of interpretations. The dilemma facing the teacher is how much support is acceptable. Does one-to-one tutoring for an extended period until a skill is achieved constitute mastery? Or must there be a time gap and a new assessment context before attainment of the learning outcome is achieved? Different interpretations lead to different demands on students, and could result in two apparently identical assessments not being comparable.

An associated question is whether the teacher teaches towards the test or towards attainment of the learning outcome. In assessing the attainment of concepts, for example, it is considered good practice to ask the student to apply a concept in a completely new situation, provided this makes no inappropriate assumptions about knowledge or context. The teacher who uses this approach will make better assessments of attainment of the concept than the one who embeds assessment in a context which is already familiar to the students.

At a more general level, it is important that when teachers are making assessments they base their judgments on actual student performance and not on their view of a student's 'general ability'. This may be less of a danger in formal or written tests, where teachers are frequently surprised by students' results, but is more of a risk in the practical context or where teachers make assessments by observation.

Attention to mastery criteria

To ensure assessments of high quality, teachers must understand what the criteria actually mean. They must relate their mastery decisions to those criteria; base their decisions on individual learning outcomes of individual pupils and not compare pupils with each other; have procedures on how to deal with what they see as 'careless mistakes'; and be clear about whether they can 'trade off' difficulties in one aspect of an outcome by recognising competence in the others. In each of these areas there is potential for differences amongst teachers which in turn lead to unreliability.

Commitment to the philosophy of the Action Plan

The final set of questions relates to the commitment of staff to the assessment philosophy which underwrote the Action Plan and which is the basis of the National Certificate. Problems may result if teachers interpret this philosophy in different ways.

First, it is clear that assessment in the National Certificate is not only for summative purposes. The emphasis on formative and diagnostic assessment has obvious educational potential but it

creates ambiguity about the status of any given assessment. Students may react differently to circumstances in which they see assessment as having a diagnostic purpose (where it is to their advantage to expose their weaknesses) compared with instances where the purpose is summative (in which case they may wish to conceal their problems).

Secondly, teachers are encouraged to make students aware of the criteria on which they will be assessed, and our earlier work indicated that students prefer to be aware of these criteria. It seems reasonable to assume that knowledge about the criteria on which one will be assessed will influence the learning strategy adopted. If criteria are not made explicit, this may reduce the value of the assessment to both student and teacher. In our earlier study, there was variation in teachers' commitment to making diagnostic assessment information available to students while there was still time to overcome identified weaknesses. This, in turn, led to differences in the opportunities students were given to display mastery. In situations where the research identifies differences in the equivalence of mastery decisions, we suspect that variation in commitment could be the explanation.

Finally, we have to explore the soundness of systems which are less 'formal' than written tests. 'Incidental' assessment which arises from the day-to-day work of the class is an attractive proposition. It lays claims to validity and has advantages of economy. But because it is less formal, it does not follow that the variables we have identified as likely to influence the quality of assessment are any less appropriate. There is no doubt that information about informal assessment is difficult to collect, but it may be important to do so.

3 Mathematics: a straightforward subject?

THE MATHEMATICS MODULES

The mathematics case-study, carried out in session 1986-87, focused on two modules taught in one further education college. Neither the college nor the modules were intended to be representative. The college was chosen because a working relationship had been established there at an earlier stage in the project. Mathematics was chosen because it was felt to be a relatively straightforward subject to assess and because it represented a basic skill with wide application.

This case study was a trial run for evaluating the quality of assessment model (see Figure 2.1). We were concerned to establish whether the model would be suitable for further investigations of assessment in the National Certificate. In the event it did prove useful and was adopted for all subsequent case studies.

Module A2 (61059 in the National Certificate catalogue) was a specialist module in mathematics which was taught by one large department within the college. Teaching materials and assessment instruments used by members of the department had been developed collaboratively. In contrast, Module G2 (61052) was a general module designed to be used in a variety of contexts. This was taught by several departments in the college, including the Construction, Foundry and Fabrication, and Business Studies departments. The teaching materials and assessment instruments and procedures used were produced independently by each department. Both of these modules have since been revised by SCOTVEC. This has implications for the detail of our findings but we believe the general principles remain relevant.

Both modules were of standard National Certificate design. In A2, five learning outcomes were specified; the first four relate clearly to the content to be covered and the fifth comprises the maintenance of a workfile. End-of-module tests were recommended for summative purposes relating to the first four outcomes although little other guidance was offered. In G2, the four outcomes comprised two related specifically to the content, one requiring a project and one the maintenance of a workfile. The recommended assessment procedures laid stress on project work although 'diagnostic' worksheets appeared to be cited as a source of summative assessment.

INITIAL HYPOTHESES

The criterion-referenced model of the National Certificate postulates that if a domain is clearly defined in advance, different instruments constructed on the basis of the definition should yield the same decisions about whether or not a student has mastered it. In National Certificate modules the domain to be assessed is defined by the learning outcomes, content specifications and performance criteria. How likely did it seem at the outset that these would be sufficiently clear to yield consistent decisions?

We assumed Mathematics to be an area of the curriculum which is more amenable than others to criterion-referenced assessment.

However, we had some reservations about the definitions offered by these descriptors. In particular, we considered that they contained very large domains. Unless the assessment instruments used were correspondingly large, only a small sample of the domain could be assessed. It would be possible, therefore, for lecturers to produce different assessment instruments which comprised items assessing the same domain but which arrived at different conclusions about a student's competence because they sampled different aspects of it. Although each of these instruments could comprise items derived adequately from the definition offered, they might result in different conclusions about student competence when they were applied.

Of course, the nature of human attainment and the limitations of assessment technology mean that 100% agreement on mastery decisions between any two assessment instruments is an unrealistic goal. What constitutes an adequate measure of agreement has to be decided in each context. In the case of formative assessment in the classroom, the teacher will make that decision on the basis of the needs of the students. In the case of summative assessment it is policy makers who decide the extent of the agreement they require.

Having studied these module descriptors, we considered that although the subject area was amenable to criterion-referenced assessment, it was appropriate to ask whether the descriptors were suitable.

QUESTIONS FOR THIS CASE STUDY

The questions we asked in this case study were:

1 To what extent does the descriptor supply enough information to ensure that different assessment instruments constructed for it will yield comparable results?

2 To the extent that there might be doubt about those decisions, what might explain any apparent inadequacy?

THE RESEARCH DESIGN

We adopted a simple research design and focused only on those learning outcomes in the two modules where the descriptor indicated the suitability of short answer questions. We confined our work to one college and, in collaboration with the teaching staff involved, constructed the alternative instruments ourselves. The mastery status of the students we tested was compared with the results from using the college's own instruments. We also gathered a set of data relating to the questions posed in our Q-model in search of an understanding of our empirical findings.

The first task was the construction of alternative instruments. One reason for choosing Mathematics for our first case-study was that short answer assessment lends itself more readily to the construction and application of an alternative assessment instrument than do some other approaches. If our results raised questions about the ability of these module descriptors to lead to alternative instruments providing similar conclusions, then there must be a greater question mark against other modules which use more complex approaches.

The alternative assessment instruments were constructed in a way to maximise their similarity with the college's own assessment instruments, as well as being derived adequately from the module descriptor. This was achieved by having each researcher-constructed instrument reviewed not only by 'outside' experts but also by the college staff who were asked to ensure that the alternative instruments were 'fair' for their students. This increased the likelihood that the two tests would lead to similar mastery decisions.

A second task involved the empirical review of the assessment instruments once they had been administered to students. We were looking at the instrument difficulty and the extent to which items seemed to be functioning in the same way. The 'operational' components of our Q-model provided the source of questions for this aspect of the design.

Finally, interviews were carried out with all staff teaching these modules in the college to illuminate the extent to which professional judgment was used and the likelihood of assessments being contaminated by 'teaching to the test'. Table 3.1 indicates the numbers of staff, teaching groups and students involved in the study.

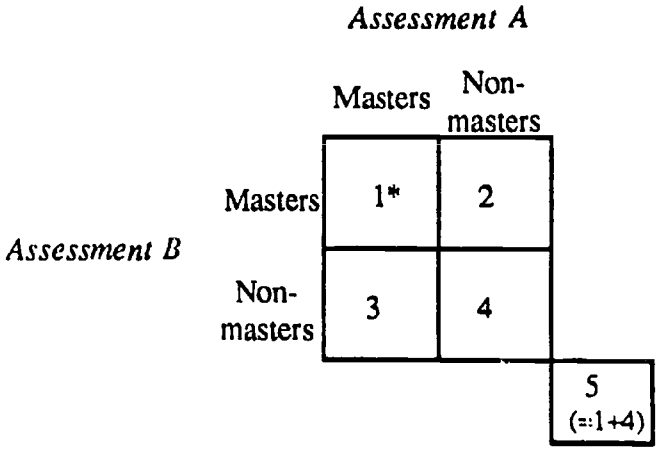
Table 3.1 Numbers Involved in the Mathematics Study

		Number of staff	Number of groups	Number of students
Module A2		6	3	31
Module G2	Department I	4	5	43
	Department II	4	4	34
	Department III	1	3	13

FINDINGS

The question being asked was whether the results from two assessment instruments would agree about whether individual students had or had not mastered the domain. In our view, the most informative and accessible approach to describe the relationship between the results is to show the raw data in the form of a simple matrix. The principle is shown in Figure 3.1.

Figure 3.1 A Key to Data Concerning the Consistency of Assessment Decisions



Percentage entries in cells 1 and 4 show consistency over the assessments. Therefore cell 5 shows total percentage consistency. Entries in cells 2 and 3 show inconsistency over the two assessments.

The module descriptor defines mastery of the module as being dependent on students achieving mastery of each of the *individual outcomes*. Figure 3.2 shows the number of students who were masters or non-masters of the *modules* in the various contexts in which the research took place. For neither module did there appear to be substantial agreement between the college assessment and the alternative assessment, although for module G2 the extent of agreement was greater. For both modules it was more difficult for a student to warrant a mastery allocation on the alternative instrument than on the college instrument.

Figure 3.2 Module Masters by Department

MODULE A2

Parallel Instrument

College instrument	Masters	Non-masters
Masters	11	81
Non-masters	0	8

19

MODULE G2

Dept I

Parallel

College	M	N-M
M	68	16
N-M	0	16

84

Dept II

Parallel

College	M	N-M
M	26	50
N-M	0	24

50

Dept III

Parallel

College	M	N-M
M	23	0
N-M	8	69

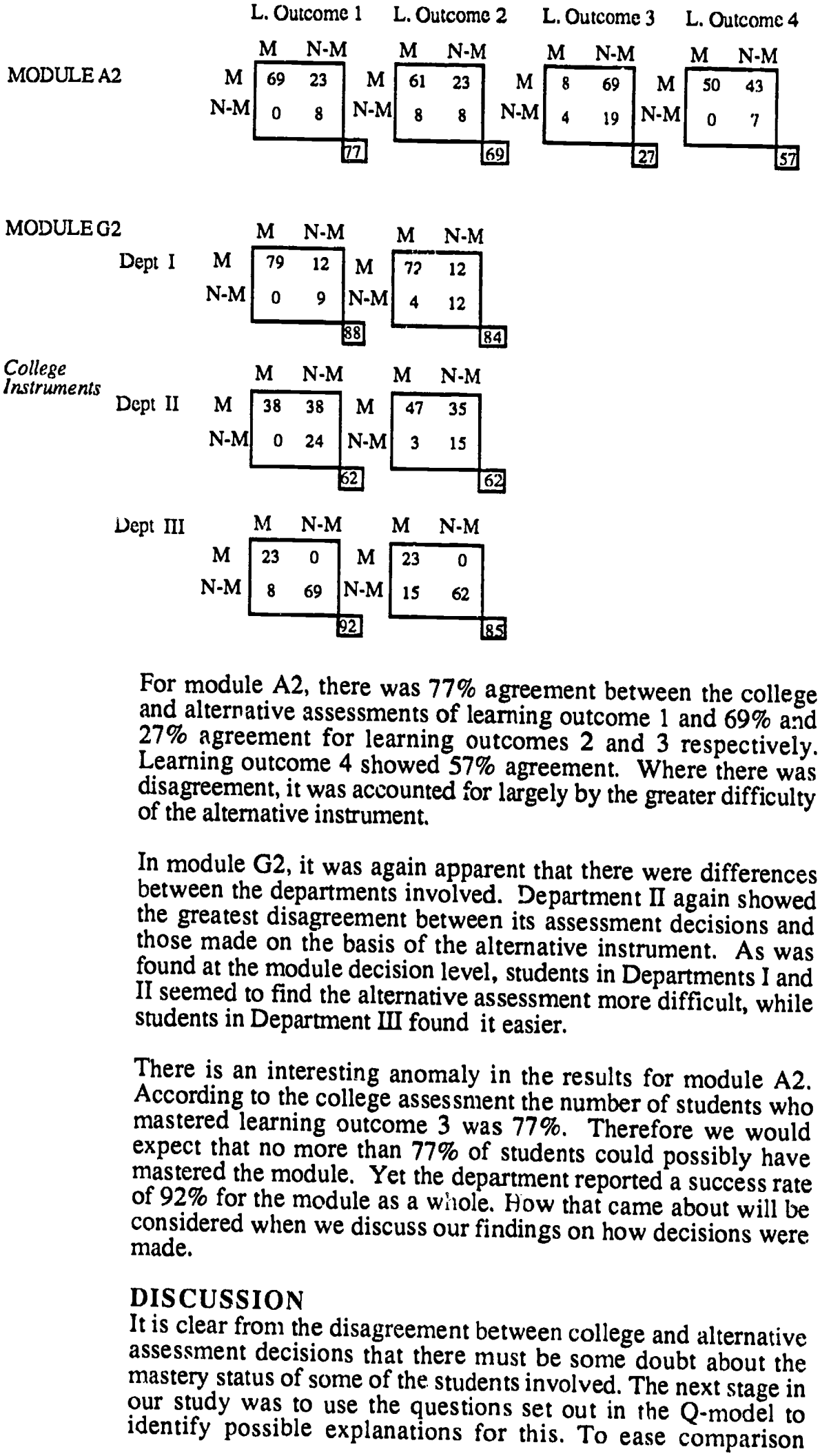
92

For module A2, at 19%, the agreement was low and the source of disagreement came entirely from the alternative instrument being more difficult than the college instrument. In module G2, there was a substantial difference between departments in the agreement between their assessment decisions and those made on the basis of the alternative instrument. For Department I, there was an 84% agreement, with the disagreement coming entirely from students who performed less well on the alternative assessment (16% of students 'passed' the college assessment but 'failed' the alternative assessment). In Department II, the agreement was 50% and again it seemed that the alternative assessment was more difficult for the students. At 92% Department III showed the greatest agreement, but in this case the alternative assessment appeared to be easier for the students than the college assessment (the 8% disagreement was caused by students who had 'failed' the college assessment, 'passing' the alternative assessment).

Mastery of the module subsumes mastery of the individual learning outcomes. Figure 3.3 therefore takes the analysis a stage further and looks at the mastery decisions for each learning outcome in each instrument.

Figure 3.3 Module Masters by Learning Outcome

Parallel Instruments



For module A2, there was 77% agreement between the college and alternative assessments of learning outcome 1 and 69% and 27% agreement for learning outcomes 2 and 3 respectively. Learning outcome 4 showed 57% agreement. Where there was disagreement, it was accounted for largely by the greater difficulty of the alternative instrument.

In module G2, it was again apparent that there were differences between the departments involved. Department II again showed the greatest disagreement between its assessment decisions and those made on the basis of the alternative instrument. As was found at the module decision level, students in Departments I and II seemed to find the alternative assessment more difficult, while students in Department III found it easier.

There is an interesting anomaly in the results for module A2. According to the college assessment the number of students who mastered learning outcome 3 was 77%. Therefore we would expect that no more than 77% of students could possibly have mastered the module. Yet the department reported a success rate of 92% for the module as a whole. How that came about will be considered when we discuss our findings on how decisions were made.

DISCUSSION

It is clear from the disagreement between college and alternative assessment decisions that there must be some doubt about the mastery status of some of the students involved. The next stage in our study was to use the questions set out in the Q-model to identify possible explanations for this. To ease comparison

between this and the other case study findings set out in this report, we will explore these findings in the order in which the model was described in Chapter 2 (see Figure 2.1) considering the 'precursor' and then the 'operational' variables related to the instrument issues before looking at the quality of the assessment process.

The Quality of the Instruments: Precursor variables

Our approach has assumed that the quality of the instruments is dependent on two sets of variables. The first, the precursor variables, comprises the set of information available to staff to guide them in their construction of instruments and in making decisions about student competence. The second, the operational set of variables, is available once the instruments have been constructed, the assessments have been made and the results are open to inspection. We will look at each of these in turn.

Adequacy of the domain definition

We suspected that the Mathematics modules domains were too large and that this was likely to pose problems for assessment, not least because the time available left little scope for long tests. One indicator of such problems came from lecturers reporting difficulty in teaching all the content of some of the learning outcomes in the modules. Another was the view of some staff that learning outcomes were not adequately defined:

> The guidelines aren't there ... You're relying on your interpretation virtually of a single sentence ... and you then open that up to all sorts of different interpretations.
>
> 'Explain something' - do you mean 'explain'? Do you mean 'show an understanding?' Do you mean 'state something'? ... it's all in the nuance of the word.

This view was shared by the research team as it considered what to include in the domain sample for the alternative instruments. Our conclusion, therefore, was that at least one cause of the low levels of agreement observed between instruments was the inadequacy of the domain definitions of what was to be assessed.

With module G2, we encountered a more specific problem. G2 is a general module which can be delivered in a variety of vocational contexts. The team had difficulty in creating an alternative assessment instrument which could be used in the very different vocational contexts of the three departments. We were not convinced that a student's possession of G2 from any one department should be construed as a comment on the student's attainment of the learning outcomes in any other context. This was a view shared by at least one member of staff:

> If somebody changed out of one specialism to another you would imagine it [G2] would be far more related but maybe ... they might need to go and take the module again.

While there are dangers in arriving at general conclusions on the basis of single case-studies, we became aware that we should ask what it is possible to assume from the possession of a 'general' Mathematics module? Does a 'G2' obtained in a specific

vocational context imply the mastery of a generalisable range of mathematical competencies? Or should we consider the mastery of these competencies as highly context-dependent? We will return to this question since it clearly has implications beyond the case-study considered here.

Clarity of what is meant by mastery

The nature of the performance criteria set out in the module descriptors varies substantially. In A2 for example, the criteria for the outcome 'the student should maintain a workfile competently' are set out in some detail. The content for a workfile is stipulated and the notion of 'competence' is tackled in terms of accuracy, skill and process. There is much left to the interpretation of the teacher (for example, what does the requirement that 'written work should be clear' mean?) but one is provided with at least the beginning of an idea of what success in 'maintaining a workfile' implies.

The current study did not consider this or other 'long term' outcomes such as those associated with project work. And in concentrating on the more traditionally assessed outcomes we found that the nature of the performance criteria supplied was different. Thus, again in A2, the performance criterion relating to outcomes 1 to 3 is that:

> The student should perform to a standard acceptable to the examiner. The exact level of performance sought will depend on the actual test set but it is likely that a score of around 75% on questions associated with *each* Learning Outcome will be appropriate.

Much is left to the discretion of the teacher. What, for example, must be included in the test? The content of the domains associated with these outcomes is substantial. How is it to be sampled? Does each aspect have equal weighting? Would a performance be classified as satisfactory if a student was unable to cope with one or more key aspects of the domain but answered sufficient of the other items correctly to score 75%? How does the teacher decide on the 'exact' level of performance required? Any problems associated with arriving at a national or even a local standard for this module will only be exacerbated unless guidance is given to staff on what is considered acceptable.

The descriptor goes on to indicate that if the student fails to meet this criterion, staff can modify their decision by reference to the workfile. However this only increases the extent to which the system is relying on an assumed shared understanding amongst staff to attain reliable assessments of student competence. It was therefore our view that the meaning of mastery of the outcomes with which we worked in these modules was less precise than it might have been, and was a potential risk to the maintenance of quality.

It was clear from the interviews that some staff also felt that the advice given in the descriptors was inadequate to decide what would constitute mastery:

> I've had a student ... he had been at another college previously and ... I found he was struggling very badly and about five or six weeks after he joined the class he brought in a certificate from SCOTVEC to say he'd already passed this module ... He's going about with a certificate which says he's reached the learning outcomes of Maths 2 and I doubt very much if he would ever have passed them in the module I was running.

This insight is an unusual one, from circumstances which occur very infrequently. It does, nevertheless, underline the importance of arriving at a shared and clear understanding of what mastery means. To imply, however, that this is a straightforward task would be wrong.

Adequacy of the instrument specification

The support given to staff on assessment procedures for the learning outcomes considered in this study was minimal. In A2, there is an indication that assessment should take place through an end of module test and that summative assessment will be based on short answer and extended answer questions. In G2, staff are encouraged to make use of work in the module project, although how this will relate to individual aspects of the domain definition is not stated. Where important topics are not covered by the project, it is suggested that 'diagnostic worksheets' be used. How *diagnostic* assessments are to be used for summative purposes is not discussed but this is a potential source of confusion for staff and student alike.

Given the complexity of assessing large domains reliably, it must be clear that the lack of discussion on the nature of appropriate instruments, on basic instrument construction techniques, and on the central question in criterion-referenced assessment of the length of test necessary to achieve reliable measurement, must be a legitimate criticism of the descriptor (as it was at the time of the study). We did not evaluate the consequences for quality of these inadequacies because we chose to mirror as closely as possible the style of assessment used by the departments in their own instruments. We would expect, however, that were we to have adopted the freedom in instrument design offered by the descriptor in constructing our alternative instruments, the extent of agreement observed would have be substantially lower than our current findings suggest.

The Quality of the Instruments: Operational variables

Adequacy of the components of the instruments

The individual items comprising the researchers' alternative instruments were subjected to review by a group of Mathematics experts and by the college teaching staff. In both cases, reviewers were asked to identify items which did not fit the domain definition offered by the module descriptor. In addition, college staff were asked to indicate whether they felt the individual items were appropriate for their students. Although the college assessment instruments were not subjected to the same rigorous review process, inspection by the research team suggested that there was little doubt that they fitted the domain definition.

Adequacy of the instruments

The good fit with the domain definition which we discovered for individual items was no guarantee of agreement in results from the

different assessment instruments. Much depends on whether individual items, taken together, constitute an instrument which is a sound sample of the whole domain. In an extreme case, if inadequate domain sampling techniques are used, assessments which appear to fit the domain definition may have little overlap in content and context.

In the construction of our alternative assessment instrument, we deliberately tried to avoid this by making it congruent with the college assessment instrument and not just the descriptor. However, analysis of our data led us to believe that there were some problems with the college instruments(and consequently our own) in this regard. The short tests used by the departments inevitably meant that there would be gaps in the domain sample.

More accurate 'domain scores' could have been computed using sophisticated statistical techniques. This would be a very difficult course to take, however, in a college-based assessment context. Another solution might be longer tests. Although this might be a good 'assessment' strategy, it could be criticised on pedagogic grounds. A clearer policy statement on domain sampling and more realistic domains would be the most appropriate starting point for improvement.

The concern expressed here was not one expressed by the case-study staff. Their priority was that the content and context of the assessment should be compatible with their students' experience. To the extent that this was achieved, they were satisfied with the assessment instrument:

> I try to base it on engineering ... and try to relate the calculations to what they would actually see in their work.
>
> You've got to apply your knowledge of the class, your knowledge that you have yourself and say, 'What can I read into that content?' knowing ... that the content is the flexible part of it, rather than the rigid outcomes, and at the same time still be saying, 'Does that meet what the outcome's asking?'

Decisions issues

When we compared the individual learning outcome and module decisions relating to A2, there was a surprising anomaly. Although only 77% of the students appeared to have mastered outcome 3 according to the end-of-module test, 92% of students were credited with having completed the module. How had this come about?

We discovered that the department concerned did not base its decisions for the module as a whole on the assessments made for the individual outcomes, but on an aggregate score over all the learning outcomes. Thus a student who failed to master learning outcome 3 could still be deemed a 'master' of the module, provided he or she did well enough on the other learning outcomes to bring the score for the whole test up to the cut-score of 75%.

Our first assumption was that this was a misinterpretation of the performance criteria stated in the module descriptor. Further

enquiry indicated that the situation was more complicated. It transpired that the department was using a set of guidelines which they had obtained from a support group in another authority. These offered advice on appropriate types of items and lengths of assessment. They suggested a mixture of formative assessment throughout the modules, and summative assessment using end-of-module tests containing items covering all the appropriate learning outcomes. But the advice on how to obtain a module mastery decision was of particular interest:

> A score of 45 or more out of 60 will be deemed satisfactory, no matter how the marks are obtained.

At least 15% of students considered by the college to have attained the module as a whole had not attained one of the learning outcomes. This runs counter to the principles of the National Certificate and indeed to the instructions given in the module descriptor. It was based, however, on apparently 'official' guidelines which the staff took on trust. The department concerned is now aware of this and has ceased to aggregate scores in this way. The problem could have been avoided from the beginning if better guidance had been available to the staff in the 'assessment instruments' section of the module descriptor itself.

This obvious source of error cannot account for all of the discrepancies apparent between the results obtained from the college and alternative assessment instruments. Indeed, the lack of equivalence in the G2 comparisons was not associated with this practice and, as far as we could ascertain, staff in these departments were following the module guidelines accurately. We suggest, therefore, that the different decisions arrived at in these cases may be explained in four ways. First, because the tests were taken at different times, they reflect genuine differences between the students' mastery states on the two occasions. Secondly, all assessments are subject to error and so small differences are expected even from closely parallel tests. Thirdly, the observed scores used are only estimates of a student's 'true' score. And finally, the practice of assessing large domains using short tests is suspect and likely to lead to a lack of equivalent assessments when different instruments are applied. The last of these sources of error would be the easiest to tackle.

The Quality of the Process of Assessment

We have suggested that the quality of assessment decisions is, in part, dependent on the adequacy of the process of assessment. We sought information on this through our interviews with staff.

There was no suggestion that the staff were antagonistic to the philosophy behind the Action Plan. Furthermore, it was a mark of their professionalism and commitment that they agreed to take part in this study. Nevertheless, one suspects that if they had been more aware of the central importance of working towards student attainment of discrete learning outcomes, those teaching module A2 would have been less inclined to aggregate scores over learning outcomes and so defeat the purpose of criterion-referenced assessment. It seems likely, however, that 'responsibility' for this did not rest solely with the staff. The haste with which the system was implemented left little time for staff to reflect on the underlying philosophy. The limited opportunities for staff development which existed, and the need to concentrate on the

mechanics of getting the system up and running, probably also contributed to the problem.

The process of assessment is closely linked with teaching. One way of examining the professionalism of teachers in assessment is to determine whether they are 'teaching to the test'. Internal continuous assessment could result in assessments that were essentially a repetition of what was taught. In this case our interview data, which were unsupported by observation of practice in the classroom, limit the inferences we can draw. However, it was our impression that, whilst some lecturers gave students assessments which were not very different from what they had been taught, the extent of this was not of great importance. The substantial majority of those interviewed clearly had a responsible attitude to assessment, preferring to prepare their students for their longer term requirements than coach them to pass the module in question.
Because the assessments for these modules were based on end-of-module tests, there was little scope for staff stereotypes of students to interfere in the mastery decision. To the extent that the test itself constituted the decision-making process, and because of the limited number of items available, we can make little comment on other aspects of the 'assessment professionalism' component of the model. The decision to base assessments on formal instruments limits the extent to which doubtful assessment practice can distort the decision-making process.

The final aspect of the process to be explored relates to the attention which staff pay to the indicators of mastery. The conceptualisation of mastery in the module descriptor was limited to cut-scores and staff were generally 'professional' in adhering strictly to them. There were, however, two notable exceptions. In borderline decisions, some lecturers used the flexibility of their marking schemes to the student's advantage, especially where a student whom they perceived as 'good' was in danger of 'failing'.

> You don't give them nothing for getting so far ... if they give the impression that they're trying their hardest I'll try to get them up to the borderline level, and it'll be touch and go whether they get through.

At least one lecturer thought that a difference of a few marks around the cut-score was not significant since the standard required was much higher than he considered necessary.

> I have students ... who are perfectly able and capable of earning 50%, maybe even 60%, and should in my opinion pass, but they'll be failed because they can't attain 75% because of the guidelines.

We do not know what levels of attainment this lecturer has in mind when he talks of '50%', '60%' and even '75%', but clearly he is not basing his decisions solely on criteria established *a priori* within the descriptor. In essence he was using his formative knowledge of previous student attainment to adjust his mastery decisions.

In the second exception to strict enforcement of the cut-score, one lecturer considered it an inappropriate way of differentiating between masters and non-masters in project work (one learning outcome of module G2). This lecturer felt that it was important

that students made no mistakes at all in this learning outcome, which was considered crucial to the module. We have no evidence on this learning outcome and so cannot offer comment on the consequences of taking this very harsh stance.

CONCLUSIONS

In this case-study, one of our main objectives was to discover if the Q-model would help us understand the factors affecting the quality of assessment in practice. It had some limitations. For example, in practice 'precursor', 'operational' and 'process' variables interact in complex ways and the same, or similar, issues crop up under different headings. However, we found that it did help us to keep in perspective the complex web of issues which criterion-referenced assessment raises. On balance, although it imposes a somewhat mechanistic style on our report, we felt this was outweighed by the help it provides in a sometimes complex analysis. We decided, therefore, to adopt it both for the rest of our case studies and as the vehicle to report our findings.

Among the precursor variables we can identify reasons for variations in the quality of assessment which have their origins in the module descriptors issued by SCOTVEC. We found it appropriate to ask questions about the breadth of the domains to be assessed, the fact that some learning outcomes are poorly defined, the weak conceptualisation of mastery in the module descriptors, and the paucity of information on instrument specifications and decision-making procedures. Some of these features of the descriptors are themselves a consequence of the decision that these Mathematics modules should be capable of being applied to different vocational contexts. A certain breadth of definition is a necessary concomitant of this decision. Other features of these descriptors - such as the weaknesses in the areas of mastery decisions and instrument specifications - were more avoidable.

Among operational and process variables we encountered aspects of practice which affected the quality of assessment. These included a lack of awareness about the consequences of sampling domains in different ways, the absence of a common framework in which staff could deal with 'borderline' decisions and general difficulties which are associated with a reliance on the professional judgment of lecturers. This will continue unless efforts are made to ensure that members of staff have a shared understanding of the demands of the module.

From the perspective of the staff involved, some of these difficulties may seem inevitable, given the inadequacies of the module descriptor (eg large domains and lack of guidance). Other problems, such as a lack of a shared understanding of the requirements of the modules, and the problems caused by the way in which staff aggregated scores were more avoidable.

We began this case-study thinking that Mathematics was a relatively straightforward subject which would provide few assessment problems. We found otherwise and wondered how other, less easily definable, subject areas were faring. Some of these will be examined later in this report.

4 Stock Control: the exchange of instruments between colleges

The second case study was carried out in Session 1987-88. It focused on the Stock Control Module (63107) and asked questions not only about the comparability of instruments designed by the same department, but also about the comparability of decisions arrived at when instruments were exchanged between departments.

THE STOCK CONTROL MODULE

The Stock Control Module is a general module which is seen as appropriate to the needs which obtain in a number of vocational contexts. There are four learning outcomes, and for three of those on which our study focused, the assessment procedure specified is a 10-item objective test. Fairly full information is given in each part of the module descriptor, and a sampling strategy is specified for each outcome identifying the balance of items required in relation to each aspect of the domain.

INITIAL HYPOTHESES

Initial inspection of the module descriptor suggested that it offered a relatively tight domain definition. This, together with a 'tight' assessment instrument mode (multiple choice), formed the basis for two hypotheses. The first was that assessment instruments derived from this module descriptor would be valid. The second predicted that the application of these different instruments would have a high probability of producing equivalent mastery decisions for students. It was this second hypothesis which formed the basis for the format of our case-study. In order to test for 'equivalent decisions', we administered alternative assessment instruments, as in the Mathematics case study, but in this instance the instruments were devised by the colleges rather than 'externally'.

QUESTIONS FOR THIS CASE-STUDY

The questions addressed included two which were similar to those in the Mathematics study and two others. In the case of the former we were interested in:

1 whether the descriptor was sufficiently tight to ensure that instruments relating to given outcomes constructed by the *same* teachers or departments would yield comparable results, and

2 to the extent that comparable assessments were not being made, whether this could be explained by aspects of the descriptor, the process of assessment, and/or the characteristics of the instruments themselves.

In the case of the latter we wanted to know:

3 whether the descriptor was sufficiently 'tight' to ensure that instruments relating to given outcomes constructed by departments in *different* colleges would yield comparable results, and,

4 to the extent that comparable assessments were not being made, whether this could be explained by aspects of the descriptor, the process of assessment, and/or the characteristics of the instruments themselves.

THE RESEARCH DESIGN

Two colleges were involved in the study. In College A, two lecturers in the Business Studies department, and in College B, eight lecturers from the Distribution Studies department, participated in the research. The students in both colleges were from a range of courses.

Table 4.1 Size of Sample

	Number of staff	Number of groups	Number of students
College A	2	1	18
College B	8	11	117

Data were collected on three sets of assessment instruments. Each had been designed by the staff of the colleges concerned. They comprised the existing instrument which each college was using for the module and an alternative instrument which College A was in the process of constructing.

To address the first research question, we compared the results from College A's original instrument with those from their alternative instrument. To address question 3, the colleges exchanged instruments and compared the results obtained from their own and those from the other college's instrument. To address questions 2 and 4, we gathered data from interviews with staff, questionnaires completed by them and others teaching the module throughout the country, and detailed analysis of the students' responses to test questions. These data related to the various aspects of the Q-model outlined in Chapter 2.

FINDINGS

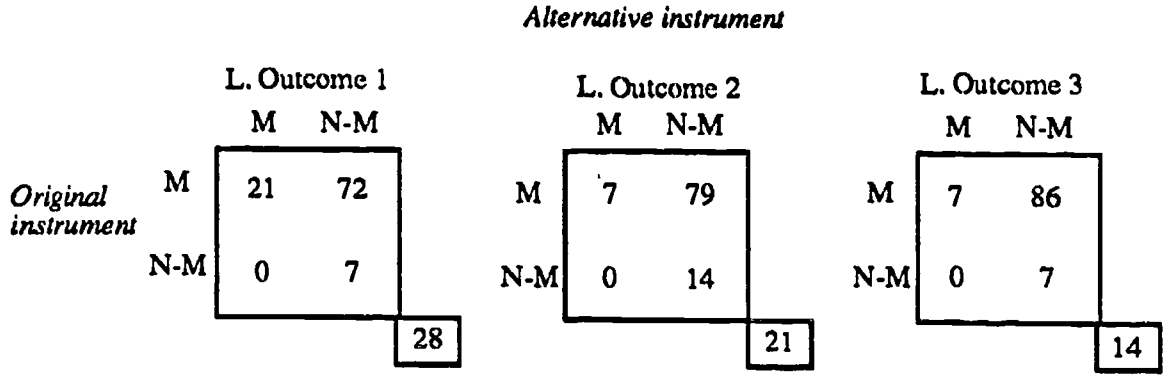
How Comparable were the two Instruments in College A?

The question being asked was whether the module descriptor was sufficiently precise so that teachers would construct instruments which would yield comparable results. Would two forms of an instrument for assessing learning outcome X, consistently agree that student Y had or had not mastered the outcome?

We show the raw data in the form of a simple matrix as explained on page 15 in Chapter 3. Figure 4.1 shows the extent of agreement between the two forms of instrument in College A for each of the three learning outcomes.

For outcome 1, 21% of the students were consistently classified as 'masters' and 7% as non-masters. In 72% of cases the instruments were not in agreement. The total agreement for outcome 2 was 21% of the decisions and for outcome 3, 14%. The 'alternative instrument' was the more difficult in each case. For learning outcome 1, 93% of students were classified as 'masters' by the original college test but only 21% by the alternative test. The equivalent figures for outcome 2 were 86% and 7%, and for

Figure 4.1 Results of Assessment Decisions for the Two Sets of Instruments for College A, for each Learning Outcome

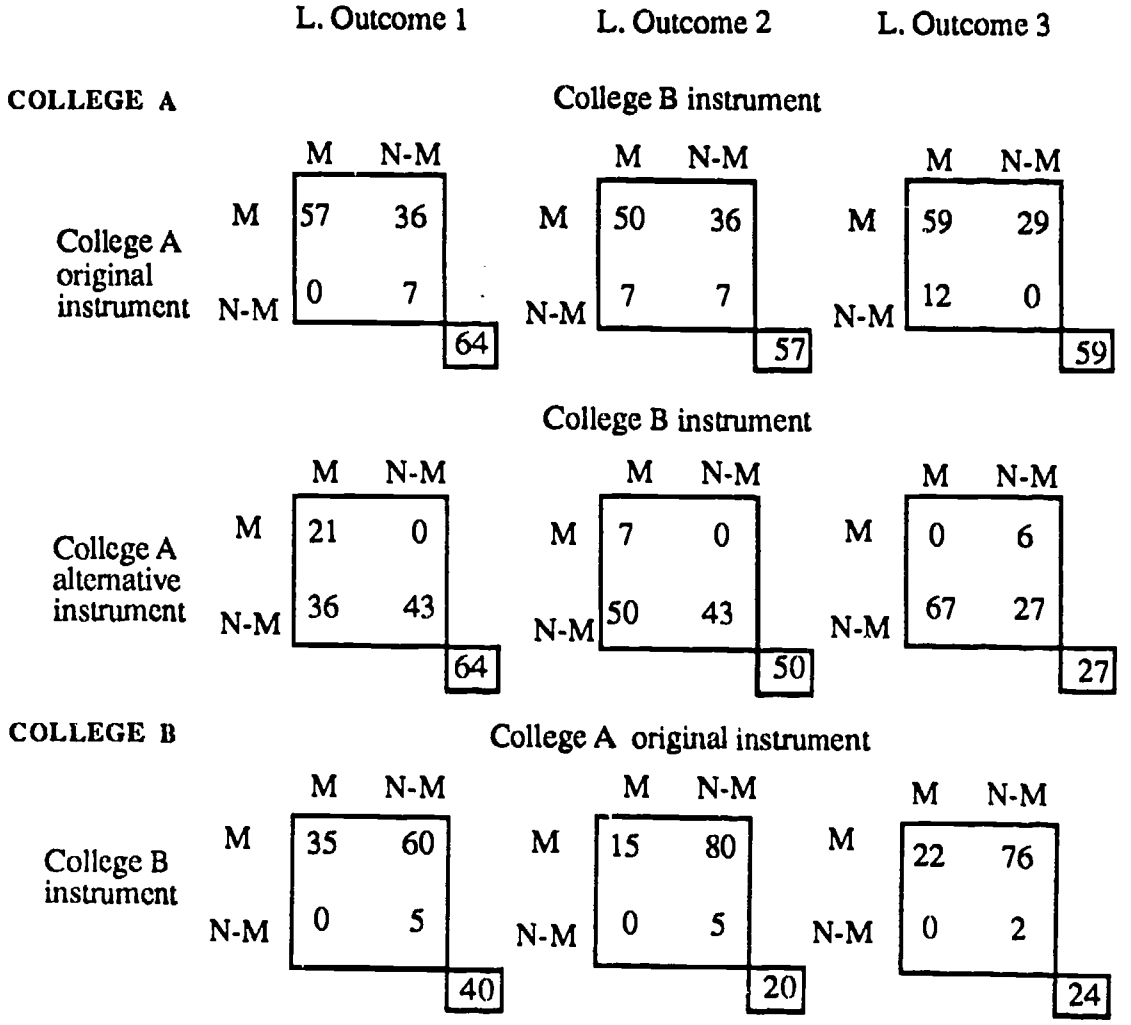


outcome 3, 93% and 7%. We will return to these data when we discuss our findings from the data gathered in relation to question 2.

What Happened when Colleges exchanged Instruments?

Research question 3 required consideration of the extent to which the module descriptor was sufficiently precise to enable departments in different colleges to produce instruments which, when exchanged, yielded decisions which were in agreement. That is, if student P was classified as a master on College A's instrument, would he or she also be classified as a master by College B's instrument? Figure 4.2 shows the proportions of consistent and inconsistent decisions for the two colleges' instruments of assessment for each learning outcome. In College

Figure 4.2 Results showing Consistency of Mastery/Non-Mastery Decisions over Different Assessment Instruments



A, a comparison between mastery decisions arising from both the original and alternative instruments is given.

For College A students, the agreement between the original college instrument and the College B instrument over the three learning outcomes ranges from 57% to 64%. Agreement between the alternative instrument and the College B instrument ranges from 27% to 64%.

When the College A instrument was applied in College B, the extent of agreement ranged from 20% to 40%. Only the original College A instrument was compared with the College B data in this case.

DISCUSSION

Two of our questions for this study (questions 1 and 3) asked whether different instruments applied to the same students would yield comparable results. The data reported above suggest that the extent of agreement was low, and in some cases very low. We had expected, however, that this module descriptor was more likely than others which we examined to be a sound source of support to staff in constructing instruments of adequate technical quality. In search of an understanding of the lack of comparability we turn once again to our model for analysing the quality of assessment (see Figure 2.1).

The Quality of the Instruments: Precursor Variables

Adequacy of the domain definition

Unless staff constructing instruments are clear about the outcomes being assessed and the content which is admissible in the assessment, it is unlikely that the instruments they construct will yield comparable results. Three sets of data describe lecturers' views on the adequacy of the domain definition in this module - questionnaires completed by the 10 case-study college staff, questionnaires completed by a national sample of 11 staff teaching the module, and interviews with the former group.

The findings suggested the college staff saw the learning outcomes as quite clearly defined, and admissible content as clearly set out. The size of domain covered by learning outcomes was also felt to be appropriate for assessment purposes. There was, however, less agreement that lecturers would interpret the domain descriptions in the same way. Data from the respondents in the national survey who taught the Stock Control module were less clear cut. There was a spread of opinion as to whether the domains were sufficiently well defined. While the majority felt that the context and content were reasonably clear, three of the eleven were of the opposite view.

The apparent anomaly in our findings from the questionnaire data is that it is claimed the descriptors are sound, yet they are expected to yield low levels of agreement. The subsequent interviews suggested this was because the lecturers felt that the descriptor offered the opportunity to contextualise outcomes to suit the needs of individual groups of students:

> One of the things I like about the modules is that you can actually change the approach, depending on the background of the students, without going outwith the module descriptor.

There was a clear consensus that although there was room for improvement in the clarity of the learning outcomes, flexibility was essential for a module of this kind:

> Personally I think we should try and gear it a bit more to what's happening to the students when they go back to their workplace.

Furthermore, one lecturer pointed out that a flexible approach had always been a feature of assessment:

> How can you say that someone who's got O-grade English has got the same as the next person when someone's done war poetry and someone's done modern poetry? ... You can't have vocational modules without having flexibility otherwise they're no longer relevant to the particular industry ... It's the standard of the assessment that has to be standard and not the content. It's got to be relevant to the background the student's in.

Several respondents felt it was important to have this kind of flexibility even if it meant that there might be a lack of clarity about what was being assessed.

> It's quite a good module. Well, it's quite an elastic module, but it took us quite a wee bit of discussion and work to get it to be quite a good module. I think it's probably inevitable that it's a wee bit sort of woolly or grey - in fact most of the modules are.

This tension between meeting the needs of individual circumstances and adhering to national guidelines provides at least a partial explanation for the lack of agreement between the two colleges' assessments. The differences between the two forms of assessment instrument constructed by the same staff in College A must have some other explanation. Further discussions with the department in fact suggested that the problem may have its roots in the alternative instrument having been created after additional information was available from SCOTVEC. This perhaps underlines the problems which can arise in relation to domain definitions.

Clarity of what is meant by mastery

Mastery in the performance criteria for this module is assessed by a straightforward 70% cut-score relating to 10 multiple choice items, and several exemplar items are provided.

The questionnaire study indicated that the performance criteria were seen to be adequate for decisions on whether students had 'passed' a test, and the level of attainment specified (70%) was seen as appropriate. On the question of inter-lecturer agreement, however, rating 'likelihood' on a seven-point scale, no respondent considered such agreement to be either 'very likely' (point 1) or 'very unlikely' (point 7). The majority fell around the middle of the scale. Findings from the national survey were similar. In other words lecturers were unsure as to the likelihood of inter-rater agreement.

Most of those interviewed had had industrial experience and, perhaps because of this, it seemed that some at least did not feel

themselves to be totally reliant on the descriptor in understanding what was meant by 'success'. It was difficult to say how much their experience contributed to this, but clearly it had some influence. For example, one lecturer, in response to a question about the role industrial experience played in recognising mastery, said,

> I don't know how you answer that. Obviously I know how I would go about it and how the firms I worked for go about it.

Another lecturer suggested that the standard of work he expected of students varied from group to group:

> I look for different levels from YTS to full-time to day-release ... I would expect somebody who'd been working say in stock control to produce materials way beyond anything that National Certificate day students would produce, and I expect them to go into it a bit deeper than the YTS.

The use of industrial experience and 'professional' expectations for different groups of students was identified in our earlier report (Black, Hall *and* Yates, 1988) as means by which staff dealt with a lack of clarity about the meaning of mastery. 'Industrial-referencing' or 'group-referencing' may or may not be appropriate for enhancing the validity of measures of student success in relation to the needs of the 'user group'. However, if such approaches are permissible, they should be made more explicit in the module descriptor. One source of the discrepancy between college instruments may be the different interpretations of what is expected of students in the workplace. Assignment of mastery as a 70% cut-score would not establish a shared understanding of what a student must *do* to 'succeed'.

Adequacy of the instrument specification

The instrument specification for this module recommends the use of multiple-choice testing for the three outcomes we considered. For each outcome, the descriptor indicates the length of test appropriate (10 items) and provides a 'prompt' on the principles of instrument design ('Each question should consist of a clearly formulated stem and four options. The options should consist of a best answer and three plausible distractors'). It also offers a sampling strategy for the test: for example, for outcome 1 it indicates the need for five 'topics' to be covered by two items each. 'Exemplar' items are also provided, although for outcomes 1 and 2 only one is given.

Analysis of both the case study colleges and national questionnaire data suggests that staff had some reservations about the adequacy of this information. The guidelines on construction were 'moderately sufficient' for most, although respondents to the national questionnaire were more critical. Staff in the case-study colleges were critical of the exemplar items, and subsequent interviews suggested that some felt the need for more than one example. Staff in the national sample were critical of the appropriateness of the multiple-choice mode of assessment, and there was evidence from interviews that staff in the case-study colleges held a similar view.

Staff claimed to be using different modes of assessment in the formative and summative contexts. Practical assignments were used in the formative context to supplement the multiple-choice for learning outcomes 1 to 3, and this assessment was considered to be 'just as important as the actual learning outcome assessment'. Multiple-choice, generally, was considered an inadequate mode of assessment by itself and best supplemented by short written answers and/or practical assignments:

> very much practical assigments because the reality of stock control ... is all practical. The fact that three out of four learning outcomes are objective tests I think discredits a lot of the potential of the module.

> Although the multiple-choice questions we use are quite relevant ... I don't think that multiple-choice questions in themselves are testing enough.

> I don't know if the material has been mastered through the objective test. I think it might have been mastered through the kind of assignments that the lecturer's putting in.

Staff in both the college and national samples were unconvinced that the instrument specification provided was adequate to ensure that different lecturers would produce similar assessment instruments. Furthermore, the research team was unconvinced that instruments of the length suggested would yield comparable measures unless the domains being assessed were more tightly specified. The freedom staff have been given to design instruments, albeit within a multiple choice mode, and their perceptions about the flexibility they had in interpreting the domain definition, made it possible for staff in different colleges to produce items relating to the same skill which were so embedded in different contents and contexts as to yield different mastery decisions.

The Quality of the Instruments: Operational Variables

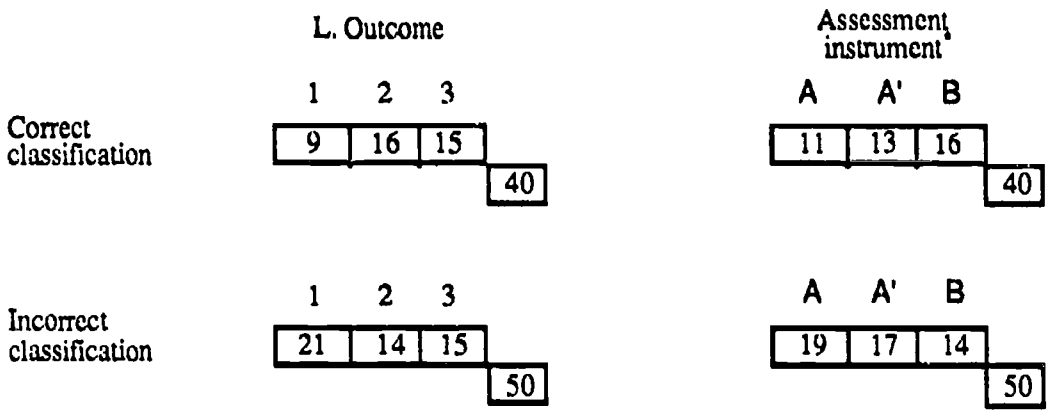
Operational variables become available for inspection only when the instruments have been constructed and/or they have been applied. Two sets of data on operational variables were collected in the Stock Control case-study. For the first of these, staff in the case-study colleges and a larger group of 18 who taught the module in other colleges (the 'evaluation group'), were asked to examine the college instruments and comment on them. This process is known as 'logical review'.

The second set of data contained information on the technical features of the instruments once they had been applied. They were subjected to a process of 'empirical review' which consisted of consideration of their power to discriminate between masters and non-masters, and of the extent to which the individual items functioned in ways similar to all the others.

Adequacy of the components of the instruments

There was a total of 90 items in the three forms of the test instrument. To consider their 'validity', in the sense that they were clearly seen to be assessing one of the learning outcomes, the 'evaluation group' was given each of these items in random order and asked to allocate them to one of the three outcomes. The results are shown in Figure 4.3.

Figure 4.3 Results for the 'Matching of Items to Learning Outcomes' Task



* A is the original instrument for College A
A' is the alternative instrument for College A
B is the original instrument for College B

Numbers shown indicate the classification of items for each learning outcome and each assessment instrument.

In the majority of cases (55%), items were allocated to an outcome which the staff who had constructed them would consider to be *wrong*. Items intended by the colleges to assess learning outcome 1 were most frequently misallocated, but the things were only slightly better for outcomes 2 and 3.

53% of the items in the instrument from College B were classified correctly, and the corresponding figures for College A were 37% and 43% for their original and alternative instruments respectively. While none of these instruments was 'satisfactory', the College B test was slightly more valid than the others and, in particular, than the original form of the College A test.

As a second stage of logical review, the evaluation group were told the outcome that each item was intended to assess and asked to consider its adequacy in more specific terms. These included its level of difficulty, the appropriateness of the content or context in which it was embedded, the extent to which it was a good assessment of the skill (or skills) contained in the domain definition, and other aspects of its technical quality. The results are summarised in Figure 4.4.

About 46% of items were considered by respondents to be suspect in relation to one or more of these criteria. Only 3 out of the 90 items were criticised on all four criteria, but 18 were criticised on more than one. Although a substantial proportion of items in each instrument was thought to have problems, the extent of these was greatest in the College A tests.

What, then, did we discover about the quality of these items from empirical review? Perhaps surprisingly, given the above, the picture was more 'satisfactory'. The question we asked about items in the empirical review process was whether each discriminated between masters and non-masters in the same way. A number of procedures were used to test the agreement between each individual item and the instrument of which it was a part.

These procedures were applied to data from a number of teaching groups. In no case did the results lead us to believe that the items

Figure 4.4 Results from the Logical Review of Items Concerning the Adequacy of Individual Items

	Number of items considered inadequate*			
	College A original instrument	College A alternative instrument	College B original instrument	Total
difficulty aspect	10	6	5	21
content/context aspect	7	6	0	13
skill aspect	9	4	0	13
technical adequacy	8	6	10	24

* where items in this category are those considered by a *majority* of respondents to be suspect in this regard

Number of items considered adequate*			
17	11	21	49

* where, for each aspect, the *majority* of respondents considered the items to be adequate

were discriminating in substantially different ways. Even if our logical review of the items and our comparison of the results of applying these instruments to different groups led us to believe that each was assessing something different, the instruments appeared to consist of items with a high degree of internal consistency. Furthermore, although some items were better discriminators than others, there was little evidence to suggest that those involved in the logical review process were particularly good at identifying them.

Adequacy of the instruments

In the final stage of our logical review, the evaluation group was asked whether they considered the instruments used to test the learning outcomes adequately sampled the domains. Unfortunately there was a poor response rate (5 out of 18) to this section of the questionnaire, but the majority of the responses were favourable, and there were few comments on anything seen to be 'missing' from the domains. The question was probably too wide-ranging at the end of a long questionnaire, and so this finding should be treated with caution.

Although individual items appeared to be discriminating appropriately in relation to the instruments as a whole, the facility values for the sets of items testing the same learning outcomes varied between 35% and 51%. This suggested that different items were not testing the same thing.

It is made explicit in the instrument specification that each of the learning outcomes comprises a number of 'sub-outcomes', and the number of items to be used for each of these is given. But what are the implications of this for item difficulty? Is it assumed that each of the sub-outcomes is likely to be of equal difficulty for groups of students taking the module? If that were the case, then it would be reflected in a 'sound' instrument comprising items of roughly equal difficulty. Or is it assumed that the various aspects of the domain are at varying levels of difficulty? In this case,

scores on items of different difficulty would be summated in order to arrive at an overall score for the instrument. No guidance on this question is offered in the descriptor.

Another variable hypothesised as significant in determining the quality of instruments is their length. There are two issues here: first, whether the specifications given are adhered to and second, whether they are adequate. In both colleges the specifications were attended to rigidly — 10 items were prescribed and 10 were used. However, we had some doubts as to the sufficiency of 10 items to test domains of the size in this module; most aspects or sub-outcomes are tested by only two or three items.

Decisions issues

The mechanics of decision-making are clearly set out in the module descriptor (70% of items must be answered correctly), and the staff in our case-study colleges adhered to this. At the same time we, and to an extent they, had some doubts as to whether this truly indicated 'mastery'.

There was no doubt that the observed score was used to make mastery decisions. As in the Mathematics study, it would seem likely that if sophisticated 'true score' procedures were applied, the mastery classifications would, in some cases, be different. Whether such procedures could or should have any place in assessments devised by teachers is debatable, but if teachers were aware of the extent to which their very short tests are unreliable measures of true scores, they might be more cautious in using the results.

The final two 'decisions issues' are 'stability' and 'equivalence'. We did not deal with the former in this case-study, and the latter was the major feature of our analysis of the consequence of exchanging instruments between colleges.

The Quality of the Process of Assessment

For some modes of assessment, the process plays a greater role in contributing to quality than in the case of multiple-choice. The clear specifications given in the module descriptor on the mechanics of arriving at mastery decisions, together with the clear-cut 'right or wrong' decisions associated with multiple-choice items, mean that decisions are based on actual student performance.

One of the process variables of assessment which could have an adverse effect on quality is the influence of the instruments on teaching: is, for example, the instrument merely a test of recall or a genuine assessment of the learning outcomes? Since we did not carry out any observation of the teaching or assessment processes, we cannot say.

However, one lecturer, a recent entrant to further education, did use the instruments as a general guide to what he should be teaching:

> I look at the question (in the assessment instrument) as well as the descriptor to see what kind of things should be covered.

In general, however, there was no evidence to suggest that staff

were 'teaching towards the test'. In both departments, there was a commitment to professionalism in the way the module was taught and assessed, and insufficient evidence on the process of assessment to explain the lack of comparability which we found when we exchanged instruments between the two colleges.

CONCLUSIONS

What did the data we gathered tell us about the four questions set out on pages 27 and 28. To begin with, our empirical studies indicated a low level of agreement between the two forms of instrument constructed by College A. This surprised us. Our initial hypothesis had been that the descriptor for this module seemed more likely than others we had examined to yield comparable instruments. Furthermore, our questionnaire and interview data suggested that staff saw the descriptor as supportive and reasonably clear. They recognised that the flexibility which allowed staff to choose content and contexts appropriate to the needs of their own students could be an impediment to comparability. But as the staff in College A presumably interpreted the flexibility uniformly, this cannot explain the lack of comparability we discovered.

There was some evidence that the technical adequacy of the original form of the College A test was suspect. Items in it were least likely to be recognised as fitting well with the domain description by the evaluation group, and fewer items were considered to be adequate in themselves. When we combine this with other characteristics of practice in College B, including the facility offered by the larger department for peer review of instruments, we suspect that an element of the low degree of comparability between the two forms of the College A instrument is the way in which the instruments were constructed. However, the more general problems of inadequate test length and the associated difficulties of domain sampling are probably equally responsible. We were more surprised by these findings than by the parallel findings in the Mathematics study.

Our second set of questions related to the comparability of decisions when the college instruments were exchanged. Again we found a low level of agreement. This can be explained partly by the factors described above, and partly by staff adapting modules to what they saw as the needs of their students.

What, then, is our overall conclusion from this case-study? First, it will not always be appropriate to assume that a student who has mastered a given outcome as interpreted by one college would necessarily be able to display mastery of it in another context. More importantly, our second conclusion is that this will come as no surprise to college staff. Comparability could probably be improved by technical adjustments to the module descriptor, but if flexibility is as important as seems to be the case, it is unlikely that comparability could be improved. Solutions to this problem are to be found in policy-making, not in testing technology.

5 Communication: too complex to expect comparability?

THE COMMUNICATION MODULES

We expected that among the modules in the catalogue some descriptors would be more 'tightly defined' than others, and that 'hard' areas of the curriculum (concerned with practical skills or strictly cognitive learning outcomes) would be conducive to 'tight' definitions, and 'soft' areas(concerned with interpersonal skills, affective characteristics of the student and other similarly elusive forms of behaviour) to 'looser' definitions. We classified the Communication modules (61001-4) at the 'softer' and 'looser' end of the spectrum, along with the Personal and Social Development modules.

It was important to look at the Communication modules because they represent essential basic skills, and are consequently very widely taught across the whole range of FE provision.

Because there are specific practical differences between the Communication modules and other modules in the catalogue we will describe them in more detail than the Mathematics and Stock Control case studies. There are four Communication modules and each has the same learning outcomes: where they differ is in their performance criteria. Each behaviour or skill can be assessed at four levels in contrast to the more straightforward 'can do/can't do' approach of most of the other modules. In this, the model of assessment in the Communication modules resembles the Standard Grade English approach with its grade-related criteria rather than the normal National Certificate approach. The existence of these four 'levels' of achievement makes Communication notably different from other modules looked at in this study.

In language studies it is common to talk of the four 'modes' of Reading, Writing, Listening and Talking. In the National Certificate Communication modules there are four *learning outcomes*, each corresponding to one of these *modes*. Each learning outcome is then sub-divided into a series of *sub-skills*. In practice, these sub-skills are rather hazily defined unless one also takes into account the additional statements about them which are included in the performance criteria for each level of performance.

So, for example, the writing mode is learning outcome 2. The actual learning outcome, as stated in the module descriptor, is that 'The student should communicate effectively in written, graphic, tabular and pictorial forms'. The sub-skills are:

2.1 convey information effectively
2.2 present ideas, opinions, arguments and judgments
2.3 describe personal experience, express feeling and reactions
2.4 employ forms of communication appropriate to purpose and audience.

In all there are twenty sub-skills covered in the Communication modules, each of which can be assessed at four levels (see Figure 5.1).

Figure 5.1 Learning Outcomes for Communication Modules

1. *Reading*
   The student should read, and where appropriate, act on communication in written, graphic, tabular and pictorial forms:
   - 1.1 gain overall impression and/or gist
   - 1.2 extract particular information
   - 1.3 recognise idea and/or feelings which are implicit
   - 1.4 evaluate the communicator's purposes, attitudes, assumptions and arguments
   - 1.5 evaluate the effectiveness of a communication.

2. *Writing*
   The student should communicate effectively in written, graphic, tabular and pictorial forms:
   - 2.1 convey information effectively
   - 2.2 present ideas, opinions, arguments and judgments
   - 2.3 describe personal experience, express feelings and reactions
   - 2.4 employ forms of communication appropriate to purpose and audience.

3. *Listening*
   The student should receive, interpret and, where appropriate, act on communication conveyed through speech and non-verbal communication:
   - 3.1 gain overall impression/or gist
   - 3.2 extract particular information
   - 3.3 recognise ideas and/or feelings which are implicit
   - 3.4 evaluate the communicator's purposes, attitudes, assumptions and arguments
   - 3.5 evaluate the effectiveness of a communication
   - 3.6 work effectively in a group.

4. *Talking*
   The student should communicate effectively through speech and non-verbal communication:
   - 4.1 convey information effectively
   - 4.2 present ideas, opinions, arguments and judgments
   - 4.3 describe personal experience, express feelings and reactions
   - 4.4 employ forms of communication appropriate to purpose and audience
   - 4.5 work effectively in a group.

While the statement of the learning outcomes is intended to clarify the nature of the domain to be assessed, the performance criteria spell out in more detail what constitutes 'mastery' for each level. This is done both by re-stating the learning outcome in fuller form and by providing performance criteria for each sub-skill.

Thus, for example, the performance criterion at level 4 for learning outcome 2, sub-skill 2 (writing: presents ideas, opinions, arguments and judgments) is stated as:

> structures and presents complex ideas and evidence in support of argument showing a capacity, where appropriate, for objectivity, generalisation, analysis, evaluation and synthesis.

Such performance criteria exist for all twenty sub-skills at each of four levels of performance. The student is supposed to demonstrate attainment of each sub-skill at the appropriate level before being awarded the learning outcome at that level. Nevertheless, at the reporting stage all that is recorded is the attainment(or not) for the learning outcome as a whole.

The Communication module descriptor does not specify the content/context within which each learning outcome and sub-skill must be achieved, but refers to a 'wide range of settings' and lists such factors as balance between productive and receptive modes as well as written and spoken modes, variety of purposes, range of settings and audience, variety of texts and different types of language use (expressive, transactional, informative etc).

In practice, the students in the case-study college (see below) were achieving these aims through the completion of a variety of 'units' of work (eg Social Dilemmas, Role Play Unit, Advertising, Job Interview Skills, Form Filling), each of which allowed them to attain certain sub-skills. These units could be based on a task or skill to be mastered (Form Filling), or could be more thematic (Social Dilemmas), and were devised by the staff of the case-study college.

The module descriptor lays great stress on the importance of formative assessment in the course of the student's work; on the maintenance of a folio of the student's work (which contains a selection of the work done in class and may be used for summative review); and on internal monitoring as a method of achieving common standards within an institution. However, it says very little about the techniques of assessment to be used and nothing about how assessment instruments are to be constructed. In this, it differs from many other module descriptors which specify a set form of assessment for particular learning outcomes. There are passing mentions of checklists and tapes as methods of recording student performance but other than that there is only the admonishment that 'the main method of assessment during the course will be the tutor's assessment of students' work in the course units'. Much depends on the professionalism which the teaching staff bring to their task.

INITIAL HYPOTHESIS

Given the 'softness' and 'looseness' of the descriptor, the range of possible content, lack of instrument specifications and complexities of the learning outcomes and performance criteria, we did not expect to find a high degree of reliability (in the sense of comparability between the levels awarded by different lecturers, or through the use of different instruments) in the assessments being made in the Communication modules.

QUESTIONS FOR THIS CASE STUDY

Assessment of the learning outcomes described above is clearly more complex than in the Mathematics and Stock Control case studies. Furthermore, assessment of each of the modes of communication in our case-study department arose from staff knowledge of student performance over a substantially longer period than the forty hours associated with most other modules. Consequently it was inappropriate to develop research questions relating to the complete assessment of a substantial proportion of the learning outcomes for the modules. Instead, we chose to focus on aspects of the assessment of three modes in some detail. In doing this we hoped to gain insights into the quality of assessments being made although clearly we must be careful not to assume that what we found can necessarily be extrapolated to other aspects of assessment in the case-study department.

The three questions we chose to ask were:

1 given the support available in the module descriptor, to what extent did staff within Communication departments and in departments in different colleges make comparable assessments of the same examples of students' *writing*?

2 given the support available in the module descriptor, to what extent was it possible for two assessors to make comparable assessments of student *talking* skills?

3 to what extent do two different approaches to assessing *listening* provide comparable decisions on student mastery?

In each case, a supplementary question was to seek to understand

the reasons for comparability, or the lack of it, through exploration of the questions set out in the Q-model.

THE RESEARCH DESIGN

The bulk of this part of the research was carried out in one department of one college during the academic session 1987-88. As part of their monitoring procedure, lecturers in this college met regularly to discuss and review the teaching and assessment of Communication. They were also involved in a Regional initiative in which representatives of all the further education colleges in the Region met regularly to monitor their provision in Communication and assist one another in developing all aspects of teaching and assessment. Through this group it was possible to involve a limited number of staff from three other colleges in parts of the research, in particular the questionnaire and writing marking exercise referred to below.

Teaching staff also completed a questionnaire and were interviewed. The questionnaire (the same as that used in the other case-studies) sought to identify aspects of the module descriptor which caused particular difficulties. After the questionnaires were completed, the staff in the case study department were interviewed in order to get a more detailed picture of particular difficulties which they had highlighted in their responses. Through the Regional group, it was also possible to have this questionnaire completed by six lecturers in three other colleges. These staff were not, however, interviewed.

Because the particular details of each of the three sub-studies are different, the design for each is provided separately below alongside our findings.

FINDINGS: 1 - THE ASSESSMENT OF STUDENT WRITING

The National Certificate module descriptors should contain all the information required to make sound assessments of students on a comparable basis. We have identified the factors which influence this as precursor variables, and we have already noted that the descriptor seemed weak in this area. We expected that there would be varying interpretations of the demands of the descriptor and that this would result in a lack of comparability between assessments made by different lecturers. In the first of our Communication sub-studies we sought to investigate the extent of comparability in assessment decisions between lecturers and to explore the reasons for it.

A sample of pieces of student writing was obtained from work done by previous years' classes in the case-study department. From these, six pieces were selected to reflect a spread of levels of achievement on a variety of sub-skills used in a range of contexts. Lecturers were asked to record, for each piece of writing, which sub-skills they thought it demonstrated and at what level of achievement. They were also asked to give some comments on the factors which influenced their assessment decisions in each case. Lecturers therefore had to assess six pieces of writing on each of four sub-skills, making twenty-four assessment decisions in all (see Figure 5.2).

The exercise was completed by six lecturers in the case-study department and six more from other colleges (two from each of

Figure 5.2: Example of Student Writing and Associated Marking Grid

Manager
Haldane Catering Supplies
15 Bathgate Road
GLENROTHES
Fife
KY2 3LS
30 January 1987

Mr J M McBride
Manager
Blackwater Cafe
119 St Kilda Crescent
KIRKCALDY
Fife
KY1 4QJ

Dear Mr McBride

I am extremely sorry for the inconvenience that we have caused you.

I will send a delivery van to collect the excess coffee as soon as possible.

The bills for the correct order will be sent to you immediately so they should already have arrived by the time you get this letter.

Yours sincerely,

Given the appropriate information about an incorrect delivery, the student was asked to write a letter of apology to a customer, as if from the manager of the supply company.

Writing Sub-skill	Level of Performance				
	N/A	1	2	3	4
2.1 convey infonnation effectively					
2.2 present ideas, opinions, arguments and judgments					
2.3 describe personal experience, express feelings and reactions					
2.4 employ forms of communication appropriate to purpose and audience					

Please give a brief outline of the factors which influenced your assessment decisions in this case:

Sub-skill 2.1:

Sub-skill 2.2:

Sub-skill 2.3:

Sub-skill 2.4:

three colleges in the same Region). We looked at whether the marking was comparable, tried to identify whether any lack of comparability could be explained by the way in which particular individuals marked, and explored the extent to which individual aspects of the test itself were responsible for problems.

Did the Lecturers mark in a Comparable way?
When we looked at the assessments made by all twelve lecturers together and applied the appropriate statistical tests* we found that they did not form a homogeneous group. However, when we looked only at the six lecturers in the case-study college, we found no significant difference amongst them, and we can accept that, as a group, they were marking reasonably comparably. Similarly, when we compared the pairs of lecturers in colleges B and C with one another we found no grounds for thinking that individuals in the pairs were marking differently. However, in college D there was a significant difference between the two lecturers from that college. Therefore, while there was consistency of marking within colleges A, B and C, there was not within college D. Nor could we say that there was consistency across all twelve lecturers.

Can we identify Lecturers whose Marking shows Significant Difference in some way?
While the statistical tests used above can indicate whether there was consistency in the group of teachers, they cannot indicate the extent to which individuals in the group compare with each other. In order to investigate this, two different approaches were used.

In the first, the awards given by lecturers were compared with two 'benchmarks': one was the levels of award which the researcher had expected each piece of writing to be given for each sub-skill; the other was the level of award given by the majority of lecturers in each case**. The two tests produced identical results. It was found that the assessment decisions made by three lecturers (lecturers 1, 9 and 11 from Colleges A, B and D respectively) differed significantly both from the levels of award expected for each piece of student writing and from the levels of award given by the majority of lecturers.

In the second approach, comparability was investigated further by testing for correlation between the levels of award given by all the lecturers***. As is shown in Table 5.1, it was found that in most cases the inter-lecturer correlations were significant and were generally high. In the cases of lecturers 1 and 9, who showed significant differences from the benchmarks on the previous tests, the significant levels of their correlations with other lecturers would lead us to suspect that they are marking reasonably

---

* The statistical test used to investigate comparability between three or more lecturers (ie. when looking at all twelve lecturers together, or when looking at the six lecturers in the case-study college) was Friedman's chi-r2 statistic. This statistic compares lecturers with each other and indicates if any differences exist, but it will not indicate the source of the difference. When looking at the colleges not in our case-study, in each of which there were only two lecturers, this statistic was inappropriate and Wilcoxon's signed rank test was used. For details of these tests see Siegel (1956).

** The 'expected' levels had been decided in advance by the researcher as part of the process of selecting pieces of writing for inclusion in the exercise. In the event, there were only two cases out of the twenty-four assessment decisions where these expected levels differed from those given by the majority of lecturers. Individual lecturers' levels of awards were tested against each of these benchmarks using the Wilcoxon signed rank test.

*** Using Spearman's rho. Again this tests lecturers' awards against one another.

Table 5.1 Correlations between Lecturers' Assessments of Student Writing

A	L2	.615											
	L3	.783	.827										
	L4	.807	.809	.687									
	L5	.846	.718	.889	.736								
	L6	.661	.746	.842	.551	.829							
B	L7	.692	.733	.828	.557	.806	.753						
	L8	.656	.643	.751	*.518*	.681	.717	.799					
C	L9	.733	.732	.628	.909	.637	*.474*	*.496*	*.469*				
	L10	.933	.616	.755	.789	.876	.664	.760	.657	.732			
D	L11	.261	.334	.189	.339	.260	.197	.106	-.039	.355	.291		
	L12	.573	.695	.529	.688	.638.	*.472*	.643	.386	.641	.674	*.457*	
	Research	.625	.973	.822.	.805	.728	.715	.739	.613	.726	.648	.399	.764
	Mode	.797	.818	.959	.685	.882	.878	.902	.789	.610	.816	.203	.591	.834
		L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	Research

Research = Research benchmark score (see text)
Correlations which fail to reach the 5% (0.407) significance level are underlined
Correlations which reach the 5% significance level but fail to reach the 1% (0.521) level are printed in *italic*.

consistently with them but may be applying a slightly higher (or lower) standard. Lecturer .1, however, failed to reach even the 5% significance level of correlation with any other lecturer except lecturer 12 (who is in the same college). On this evidence, lecturer 11 is the only one whose levels of award can be said to show a lack of comparability with the other lecturers who completed the exercise.

To what Extent could Variations in the Levels of Award be ascribed to Particular Items?

Even if we can say that, overall, most lecturers' marks correlate well with those of the other lecturers who took part in the Writing exercise, and show no significant difference in the levels of award they gave, there were variations in the marking of individual items which remained to be accounted for Detailed analysis of the lecturers' comments on factors which had influenced their decisions suggested that there were two kinds of marking disagreement. They comprised:

1 disagreement about whether a sub-skill is appropriate to a piece of writing. Unfortunately, when lecturers thought that a sub-skill was not applicable they tended not to give reasons for their decisions.

2 in instances where there was agreement about the skill involved, disagreement about the level of award appropriate to the piece of writing. There were several examples where the awards spread over three levels.

As an example of the complexity of the domain being assessed, an examination of the information given in the learning outcome and performance criteria for sub-skill 2.1 (Writing: 'conveying

information') suggests that there are at least *eight* dimensions or constituent elements by which a piece of student writing is to be judged, six in the restated learning outcome and a further two in the performance criterion (and this is only one out of twenty sub-skills). These are listed in the table below along with an analysis of the number of occasions on which lecturers chose to use them as reasons for awarding a given level of competence for one of the pieces of student writing.

Table 5.2 Constituent Elements of a Performance Criterion

Element	Times cited
Complexity of information	1
Range of media (written, graphic etc)	0
Ease of comprehension	2
Formal features (structure, tone etc)	5
Appropriateness (purpose and audience)	1
Context/situation	0
Clarity of communication	4
Completeness of information	3

There appears to be a number of different features of a communication of which lecturers are asked to take account, and it is clear that in this case study they chose to focus on different elements when justifying their awards — even in cases where the decisions they arrived at were comparable.

FINDINGS: 2 - OBSERVATION AND PARALLEL ASSESSMENT OF TALKING

Many of the same questions recur in the second sub-study which considers the assessment of Talking, with the added complication that in this context we are dealing with relatively intangible skills which are not easy to record and which often require 'on-the-spot' judgments to be made by the lecturer. We wanted to investigate whether two assessors could make comparable judgments of student performance in Talking. We also hoped to investigate further the process by which observations are made, recorded, and turned into judgments of student performance by reference to the performance criteria.

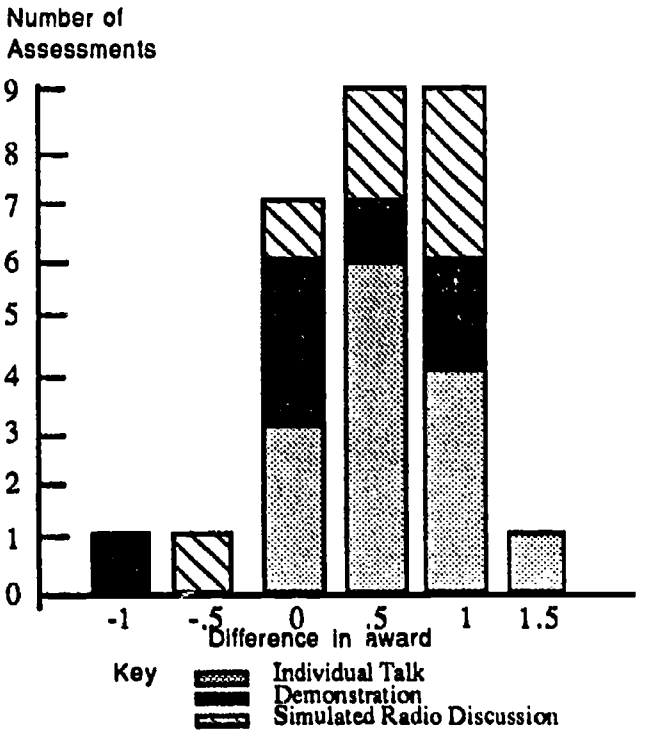
The researcher followed a group of Media Studies students through two classes for four weeks. In that time the students were conducting simulated radio discussions in one class and giving individual talks and/or demonstrations in another. Both of the lecturers concerned stressed that this was very much formative assessment and, since it was also the first time that the students had attempted the exercises upon which they were to be assessed, the results were not likely to reflect their final performance at the end of the year. The researcher had a background in English, was familiar with the Communication modules, and had had no previous contact with this group of students.

A Comparison of Assessment Decisions

In all, the researcher made 28 parallel assessments which he was able to compare with those of the two lecturers involved. It was necessary in making comparisons to introduce the notion of 'half levels' to take account of examples which were reported as '3-' or '-3' or '2/3'. These would each count as two and a half. The differences between the levels awarded by the researcher and the

lecturers are shown in Figure 5.3. In almost all cases the lecturers awarded a slightly higher level of performance than the researcher.

Figure 5.3 Differences in Award Between Assessors in the Talking Task



The degree of consensus apparent is quite high. 17 out of 28 decisions were within half a level and 27 out of 28 within one level.

Making the Assessments

The first assessment that the researcher observed was one of the simulated radio discussions. At first the checklist was used as a straightforward assessment grid which was completed (by putting a tick against those items achieved by the student and a cross where some deficiency was evident) as the exercise proceeded. This soon became unsatisfactory. Many of the behaviours being observed were too complex to allow for this sort of instant analysis into constituent parts. Nor was it always clear whether a particular behaviour was more properly categorised under one heading or another. Some apparently important events did not clearly fit in anywhere. Also, one occurrence of a particular behaviour could often be contradicted (or one's judgment upon it modified) by some other behaviour. It became apparent that what was needed was a combination of detailed notes on particular points and some form of synthesis or overview. Therefore the checklist was kept to one side and used as an *aide memoire* of possible behaviours, skills, etc to observe. Most time was spent in writing loose, and often hurried, notes in an attempt to capture the important aspects of the performance. Observation of the lecturer, and subsequent conversation, confirmed that this was also the lecturer's practice, as it was for the other lecturer involved with this class. Given the complexity of communication skills it is perhaps not surprising that this should be so. All further observations were conducted in this way.

As soon after each session as possible, and while the events were still fresh, these notes were written up by the researcher and, in the light of the performance criteria, a level of award was attached to each performance. At the end of the four-week observation period the lecturers' assessments of the students were obtained, as were copies of the notes they had taken during the classes.

An example of the assessments made by the researcher and the lecturer of a student talking about Drinking and Driving is given in Figure 5.4. While this is only one example from the twenty-eight assessments made, it does raise points which were common to many of the assessments.

Box A: The notes taken by the researcher during obervation of the student's talk, in their subsequently re-written form.

Box B: The levels of award for each sub-skill, together with reasons, as awarded by the researcher.

Figure 5.4 An Example from the Observation and Assessment of Talking

A	C
Researcher Notes A very uncertain and nervous performance. So 'laid back' that he was almost hiding under the table - clearly extremely nervous. The introduction was marred by hesitation and immediately followed by an open question to N (known to be garrulous and sure to say something). The discussion was then thrown open by asking 'Does anyone else agree?'. The idea of drinking and driving as the 'most socially accepted crime' was somehow introduced here but was neither explained nor elaborated on. There did not seem to be any links between the questions that D was asking - they just appeared to be getting fired off one after another from a prepared list and the response to others contributions was sometimes sharp - as in one case where it consisted of an abrupt 'Why's that?'. At one point the discussion degenerated into something of a melee and at another N and C hogged the floor - there was little evidence of attempts to control the discussion in either case (no attempt to calm group or to bring others in). A question mark had to be raised over the whole shape and continuity of the discussion - D didn't seem to know where it was going. Issues were raised by the group but were not properly explored - J raised the issue of how blood alcohol levels were set and C pursued this into the notion that drink affects different people differently - then N continued with this theme. D suddenly introduced some opinion poll results - without explanation - and a period of general denigration of the government followed. All very disjointed. There was also some factual confusion about how breathalysers work and blood alcohol levels are determined, but the students either didn't know about the difference or didn't want to say. Winding up of discussion, D introduced comments about a TV programme on road deaths - wouldn't this have made a better introduction? S introduced the idea of the car as a lethal weapon, but too late! It's all over.	Lecturer Notes Not an adequate introduction Not confident at beginning Some questions gave yes/no Too many questions, too quickly Gaps - not enough response Eye contact limited to speakers, not whole group Use of hand - good - limited interviewees [against 'controlling interview' item] Little comment on contributions [from interviewees] - Register/Colloqualism - Tone towards N - rather accusing 'Agree with me' - keep yourself out of it Yes answer - probe further Feedback - do not nod - verbalise it C - took over the commentary in places Links - 'Right - Well!' - before giving statistics Bleep word - Read piece at end. Reaction to my signal too abrupt.
B	D
4.1 - 2 Voice and delivery not generally appropriate. Therefore not level 3. On the other hand no support given therefore more than level 2? 4.2 - 3? If feeling generous. Doubt about the extent to which objectivity and generalisation achieved. 4.3 - n/a 4.4 - 2 Awareness there but not generally appropriate. 4.5 - 2 Listening and questioning doubtful, as is modifying opinion. Lack of control of discussion - fit in here? Overall - level 2	(4.2) Intro poor etc (4.4) Delivery disjointed/Links not always clear Control difficult. Restricted eye contact. (3.2) (3.3) Not enough response. Level 3-

Box C: The notes taken by the lecturer during observation of the student's talk.

Box D: Notes on sub-skills and level of award for the learning outcome given by the lecturer.

It is clear that neither the researcher nor the lecturer stuck rigidly to the checklist. Instead, both preferred to note down behaviours or incidents which they considered relevant or important. This may in part explain why the researcher and the lecturer chose to concentrate on different sets of sub-skills. A more systematic use of the checklist would have required both to assess the same sub-skills. Whether this would have increased the comparability of their assessments cannot be gleaned from these data.

Both the researcher and the lecturer noted the lack of control over the discussion which was exhibited by the student. However, both attributed this to different sub-skills. The researcher decided (with some misgivings) that this was pertinent to sub-skill 4.5 (group skills) while the lecturer linked it with delivery and eye-contact and recorded it under sub-skill 4.4. Perhaps we cannot expect real communications to fit neatly into abstract schemes of sub-skills?

In addition, the researcher had difficulty on several occasions in using the words used in the descriptor to identify different levels of performance. For example, the researcher had doubts about a level 3 award for sub-skill 4.2. What are 'objectivity and generalisation' and how much is 'some measure'?

Another feature of the information given in the module descriptor which posed problems was the multiplicity of component parts of sub-skills. In sub-skill 4.1, for example, the student did achieve the level 3 criterion of communicating information 'clearly without omission of essential content' but not the demand that 'voice, content and delivery are generally appropriate to purpose, situation and audience'. However, level 2 seemed a little hard because the student should be 'given some support' and this student was not given such support. How to reconcile all these sub-criteria which exist within the performance criteria struck us as a problem which has clear parallels in the large-scale domains encountered in the Mathematics study; and in that context we clearly identified it as a potential impediment to high-quality assessment.

A similar problem emerges when we consider how the researcher and the lecturer could have arrived at their final awards for the learning outcome as a whole. How do we aggregate separate sub-skill awards? Is it valid to do so if these are really separate skills? Are sub-skills weighted one against another?

FINDINGS: 3 - THE ASSESSMENT OF LISTENING

Context is an important, and relatively unspecified, factor in the assessment of Communication skills. Assessment of listening skills generally takes place in an informal setting, perhaps in the course of group discussions. In our third sub-study we wanted to see if an alternative form of assessment would produce comparable results to those already obtained by the college.

A fairly 'traditional' type of comprehension exercise was prepared in collaboration with one of the lecturers in the case-study college.

The test consisted of a taped passage on the early life of Alexander Fleming, partly narrated and partly dramatised, and a series of comprehension questions designed to assess four of the six listening sub-skills. This research test was taken by 64 students. Two of these were discounted as they had missed part of the tape. Existing college awards for Listening were also collected for 35 of these students as a means of comparison with normal college practice.

Marking
The students' responses were marked by the researcher in three ways:

1 An overall learning outcome award was made on the basis of the entire script. For this, the performance criteria for the various levels were used as descriptions of 'typical' performance at each level to build up an overall picture of what a 'typical' level 4 student, for example, might produce. There was no attempt to check the student scripts against each individual performance criterion, but rather a 'general impression' was formed. This procedure was inevitably somewhat subjective, although guided by the performance criteria.

2 A more detailed set of awards was given for each sub-skill. In this exercise there was an attempt to adhere more closely to the demands laid down in the performance criteria. Since students often gave additional information in their answers to questions it was decided that the evidence for an award in a particular sub-skill would come from anywhere in the students' scripts, and not necessarily from questions which had been specifically designed to assess that sub-skill.

3 Questions were individually marked on the basis of a prepared checklist. The items from this checklist were then aggregated into sub-skill groupings.

These three types of marking will be referred to as 'learning outcome', 'sub-skill' and 'checklist' awards. The sub-skill and checklist awards covered sub-skills 3.1, 3.2, 3.3 and 3.5. The exercise did not adequately cover sub-skill 3.4, while 3.6 was inappropriate for the research design being used. The sub-skills were listed in Figure 5.1 on page 39.

How did the Research Learning Outcome Awards compare with the College Learning Outcome Awards?
The levels of award for the learning outcome given by the college can be checked against those given on the basis of the research test. Using Spearman's rho this shows a correlation of 0.70 ($p<0.01$). This is well beyond the correlation which could have arisen by chance and suggests a degree of agreement. It is also worth noting that almost all the observations fall within half a level of agreement.

Did the Comparisons stand up to the more Detailed Scrutiny of the Sub-skill Scores?
We have no records of the awards that the college gave these students on the individual sub-skills which go to make up the Listening learning outcome. However, we can examine the

correlations between the research sub-skill awards and the college (and research) learning outcome awards.

Table 5.3 Correlations between Sub-skills and Learning Outcome Awards

	Research sub-skills			
	sub1	*sub2*	*sub3*	*sub5*
College LO	0.24 (ns)	0.41 (p<0.05)	0.52 (p<0.01)	0.40 (p<0.05)
Research LO	0.55 (p<0.01)	0.74 (p<0.01)	0.80 (p<0.01)	0.67 (p<0.01)

As Table 5.3 indicates, the individual sub-skills show higher correlations with the research learning outcome awards than with the college learning outcome awards (perhaps to be expected since this is the same marker).

The correlations between the college learning outcome awards and sub-skill awards is noticeably lower than that between the college learning outcome awards and the overall research learning outcome awards of 0.70 reported in the previous section. In one case (sub-skill 1) the correlation is not statistically significant. This could suggest that at this level of detail the research sub-skill awards and the college learning outcome awards are beginning to diverge — that is, that they are assessing different things. However, it is difficult to come to firm conclusions as we are comparing detailed research awards at the level of the sub-skills with more general learning outcome awards from the college.

What Happens when we get to the Detail of the Checklist?

The same exercise as above can be performed using the checklist scores instead of the research sub-skill scores. Table 5.4 shows that *none* of these correlations, with the exception of that between sub-skill 2 and the research learning outcome award, is significant. Sub-skill 2 is the most concrete of the sub-skills ('extracting facts and information') and may be the most amenable to checklist-type assessment. Whatever is happening, it is clear that the checklist assessments are recording something different from the other forms of assessment. This raises the possibility that awards at the learning outcome level are based on some form of general impression which is more than the sum of the parts of

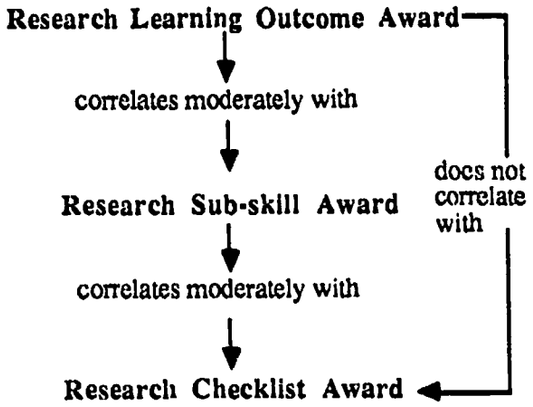
Table 5.4 Correlations between Checklist and Learning Outcome Awards

	Research Checklist			
	sub1	*sub2*	*sub3*	*sub4*
College LO	-0.14 (ns)	0.11 (ns)	-0.16 (ns)	-0.09 (ns)
Research LO	0.14 (ns)	0.49 (p<0.01)	0.29 (ns)	0.24 (ns)

either the individual sub-skills or the checklist items. The question then is whether this 'extra' element reflects aspects of communication which are lost in the analysis of communication into constituent sub-skills or checklist items, or whether it is merely 'noise' contaminating the assessment.

A final check was made amongst the three types of research award to complete the circle from learning outcome awards through sub-skill awards to checklist scores and back again. This is shown in Figure 5.5. It would appear that, while the learning outcome awards correlate with the sub-skill awards, and the sub-skill awards correlate with the checklist awards, there is no significant relationship between the learning outcome awards and the checklist scores. The more detailed the marking of the student scripts, the less likely it is to show any significant relationship with the overall award given for the learning outcome.

Figure 5.5 Relationship between Research Awards



DISCUSSION

Our initial expectation was of low reliability in the assessment of Communication. However, in some of our findings from the Writing, Talking and Listening studies there was greater agreement between lecturers and the research assessments than expected.

However, in all of these exercises, problems became apparent as we progressed to greater levels of detail. Examples from the Writing and Talking exercises showed how the complexity of the learning outcomes and performance criteria led to differences of interpretation amongst the lecturers, or between the lecturer and the researcher. Attempts at a detailed assessment of Listening skills showed progressively less correlation with the learning outcome awards as the assessment became more detailed.

Our research questions, however, called not only for an account of the quality of assessments being made but also for an investigation of the possible reasons for it. In the first instance we turn to the precursor variables and report the insights into these variables gleaned from the staff interviews and questionnaires.

The Quality of the Instruments: Precursor Variables

Some of the responses to the questionnaire and interview questions showed a diversity from which it is difficult to draw any conclusions - other than that lecturers disagree. This was the case, for example, when they were asked about the clarity of the content specifications in the descriptor, or the likelihood of there being agreement between lecturers on the interpretation of the standard

required at various levels. In other cases, questions produced much clearer responses.

Adequacy of the domain definition

One of the most obvious points about the Communication module descriptor is its sheer size, much of which is made up of extensive lists of performance criteria at each of the four levels. This complexity was evident in some of the comments which suggested that the learning outcomes were not seen as self-explanatory:

> I think it's particularly difficult in the Communication area. I think that's the one glaring example of a descriptor which is extremely complicated, which isn't easy to understand at first reading and needs a lot of back-up help. Other descriptors in other areas are easier to understand.

Although there are some slight variations between learning outcomes, it is clear that these lecturers regarded their own experience and professionalism as the best guide to what the Communication module is actually looking for. This was the first of many times in the questionnaire and interview responses when lecturers expressed a worry about interpretations varying according to the 'background of the lecturer' or the 'knowledge or skills or, in fact, experience' of those teaching the module. Even leaving aside that matter, it was admitted by one lecturer that

> It's taken us three years to really come to some sort of understanding about what the outcomes are asking for.

The problem of 'what the outcomes are asking for' is, of course, further compounded by the context in which Communication is being taught, in particular the nature and needs of the student group:

> ... a lot of the criteria that are required with totally different types of courses are tremendously varied. Anything that's demanded from Drama students just cannot be seen to bear any resemblance to anything that a catering student might need to know.

In cases where there was less concern about the varying demands of different student groups as such, there were still questions raised about the inclusion of apparently different types of content within the same learning outcome or sub-skill:

> How can you compare something in a reading area where you're maybe dealing with statistics and graphs, whereas you're taking on another occasion a very complicated article from a quality newspaper? How do you compare the level of ability?

Part of the problem, of course, is the nature of Communication itself for, as one lecturer put it:

> There's no limit to the amount of, or the variety of tasks that you can ask of a student in Communication.

These comments all seem to suggest that the domain definitions in the Communication module are far from clearly stated, a suspicion

which was borne out in some parts of the Writing exercise where lecturers disagreed amongst themselves as to whether a particular sub-skill was appropriate for a piece of student writing or not. This problem was also apparent in the Talking observation where the researcher and the lecturer disagreed as to which sub-skill a particular observed behaviour belonged.

Clarity of what is meant by mastery

Performance criteria, 'standards' and decisions about whether a student had, or had not, demonstrated 'mastery' emerged as a major area of concern for the teaching staff. Thus, for example, in reply to a questionnaire item on whether 'the performance criteria in the descriptor make it clear how to decide whether or not a student has been successful', 10 out of the 13 lecturers felt that the performance criteria are difficult to use because they do not give enough guidance.

In the responses to other questionnaire items, lecturers did not think that the assessment procedures laid down in the module descriptor would provide any guarantee that students had indeed mastered the learning outcomes.

The ambiguity and lack of precision in the performance criteria were commented on by all the lecturers in the interview. One noted that they were full of 'weasel words' which 'preclude understanding'. This was echoed in other comments:

> One of the things which people found difficult when they first got the module descriptor were these levels of input on reading, for example, where they say 'Use a text which is generally accessible' or one of 'some complexity'. Now who actually determines what is 'generally accessible' and 'of some complexity'? We have had very little guidance on that.

While this vagueness in the performance criteria did cause problems, it did not follow that all of these lecturers would have wished them to have been more tightly defined, even if they thought that were possible. In some cases there was a fear expressed of the stultifying effects of defining performance criteria too closely:

> I'm finding more and more that the assessments I'm writing are written for the performance criteria. In fact, very often I'm using the words to make sure that I'm covering the performance criteria. And I don't like that aspect of it at all. I think they're assessment-led, if you like, but I'm choosing the things which are going to assess what I've been told I must assess rather than an assessment which I think would be useful.

Another lecturer commented that assessment procedures which were too rigid would 'interfere with the freedom of individual lecturers to tailor the module to meet individual student needs'.

If we are to judge from what the lecturers said in their questionnaire responses and interviews, then we would have to conclude that the performance criteria — the bases for 'mastery' decisions — in the module are far from clearly defined. However, many lecturers feel this may be an inevitable feature of their

subject and, on balance, they prefer this to what they see as the negative effects of defining them too closely.

Adequacy of the instrument specification

In a sense, it was unfair to include questions on the adequacy of instrument specifications in the questionnaire and interview for Communication lecturers as the module descriptor does not contain any detailed specifications for them. Some of the lecturers commented on this:

> I looked at the descriptor again and I can't actually see much guidance.
>
> I didn't feel that there was enough in the module descriptor itself. Any information on instruments has come separate from the descriptor.
>
> Apart from advising that the folio of assessment should contain work which covers the full range of learning outcomes and sub-skills, there is no specific guidance on the construction of assessment instruments.

In practice, most of the lecturers (7 out of 13) ignored this question. However there were relevant comments made in other parts of the questionnaire, and in the interviews. Some of these concerned the technical difficulties of constructing assessment instruments:

> ... even something like constructing a reading test, there is a skill involved in that and often — I mean obviously when we construct a reading test we go back to the performance criteria and make sure we build in questions that allow the students to cover the criteria at the particular level that we think they are capable of. But there is quite a bit of skill in that and not all staff are actually able to do that without a bit of help. I wouldn't even say that myself. We make mistakes. Each time we use it we say 'Ah! that wasn't worded right' and you change it for the next time.

Another lecturer commented that 'it takes a great deal of time to analyse what is required before making up the assessments'. But again there was a feeling among some lecturers that it would be undesirable to have too much prescription of the assessment instruments:

> In some [modules] it's very specific, but then, if they said you must do such-and-such I would feel that that was inhibiting and that, you know, there are other ways of doing it. Perhaps it's an area where there are no solutions because the level of subjectivity is so great.

Many of the assessments made in Communication do not, of course, require 'instruments' but are based on observation of the students (Talking and Listening) or judgments made on their work in the folio. For these, it is usual to use some form of checklist to record student performance. These checklists 'translate' the demands of the performance criteria into terms appropriate to particular contexts, though it was recognised that this 'translation' had its dangers:

> ... we devised individual checklists for different styles of writing, different styles of talking and so on, but when we wrote the checklists we always kept the performance criteria in mind so that we know that what we were looking for actually referred back to 2.1, 2.4, whatever it was. They arose out of the general statements made in the performance criteria ... that worried [us] at one point — where do you draw it all back and check that you are using the performance criteria as they were originally stated?

Not all of the lecturers were entirely happy with the results:

> Assessments tend to be related to one task rather than an overall view of the students' capabilities — ie it is 'bitty' ... This may be because many of our assessment instruments are very detailed and perhaps over-complicated.

Nor was there great confidence amongst the lecturers that the assessment used would provide any guarantee that the students would demonstrate retention of the skills they were assessed on. In the questionnaire responses, only four out of thirteen lecturers expressed the view that the assessment guaranteed long-term mastery on the part of the students. However, this was qualified in some comments by the suggestion that this was probably true of all exams; that this system was preferable to an end-of-year exam; and that a guarantee of long-term mastery depended more on the amount of teaching that a student had had than on the assessment procedures:

> All it means, and fair enough, it's true of all modules, is that the person has done certain things at certain times. Having said that I think it's better than a once a-year-exam.

> If your module is only on a thirteen week block they're not having a great deal of time to master a lot of the skills, so if you're having it spread over year the chances are that the skills will be mastered in a more long-term way.

The Quality of the Instruments: Operational Variables

The operational variables - that is, what actually happens when assessment instruments are constructed and used - have, to a large extent, already been discussed in the course of reporting our findings in the Writing, Talking and Listening exercises. All that we will attempt here is a brief summary of our conclusions as they relate to the model.

Adequacy of the components of the instruments

In both the Writing and Talking exercises we noted some difficulties in deciding whether a particular sub-skill had been mastered by students, and cases where the same observed behaviour was assigned to different sub-skills by different assessors. Within the Listening exercise, the lack of correlation between the overall learning outcome awards and the more detailed assessments made on the basis of the checklist suggest that there might be some difficulty in defining the component parts of a communication in such a way as to produce accurate assessments.

It was much more difficult to decide the extent to which individual components of the instruments were adequately differentiating between masters and non-masters. While there are relatively economic techniques available to evaluate individual components of assessments in multiple-choice testing, the procedures associated with the kinds of assessment used in the Communication case-studies are both underdeveloped and time-consuming. It is difficult to judge the quality of individual assessments of talking and writing.

Adequacy of the instruments

The difficulty with the checklist scores in the Listening exercise is open to a different interpretation. It may be that the individual checklist items were sound enough but that they failed to take full account of all the required features of the students' answers. In other words, they failed to sample the domain adequately. The mismatch between the checklist scores and the overall learning outcome awards made on an 'impression' basis could then be explained by saying that the overall award is more efficient in sampling the whole domain because it is based on the professional expertise of the lecturer. Alternatively it could be that staff were taking into account aspects of performance not related to the listening exercise. We do not know which of these is the more likely explanation.

In brief, therefore, the problems of evaluating individual components of instruments outlined above relate equally to 'whole instrument' issues. They are compounded by the difficulty of knowing whether the domain is adequately sampled. There is no easy 'testing technology' solution to this lack of knowledge, but an appropriate starting point might be to heighten staff awareness of the questions to ask.

Decisions issues

In the Writing and Talking exercises, and at the learning outcome level in the Listening exercise, there was a degree of agreement about levels of award, even if there was disagreement about details. This would suggest substantial professional consensus among the lecturers. In order to investigate this further we must turn to what lecturers had to say about the process of assessing Communication skills.

The Quality of the Process of Assessment

Given that the lecturers felt there was no clear domain definition in Communication (as defined by the elements of the module descriptor), and that the basis for making decisions about 'mastery' (as defined by the performance criteria) was equally unclear, why was there agreement amongst lecturers, and between the lecturers and the researchers, on the assessment decisions they made?

One answer to this must relate to the professional 'craft knowledge' brought to the assessment of Communication by the staff involved, and seen by the lecturers as a vital factor in interpreting the descriptors:

> We have based a lot of content on what we've been doing previously ... building on experience that we already had and perhaps varying the teaching approaches and perhaps giving a bit more variety in the area of listening in particular.
>
> ... it all comes back to what we did in the past. And I think that because we're a college where particular standards were achieved and we knew what, let's say, an ONC level was we said 'Right, you know level X, that's what it is'.
>
> I think again I'm relying on past experience ... I'm not specifically always just sticking to the performance criteria.
>
> Teachers tend to refer back to what they've taught already, and if you're talking to people who've taught English already, they've therefore sort of a general consensus because they're all referring back.

It is worth noting both the frequency of references to it by lecturing staff, and their worry that inexperienced staff would have much greater difficulties in teaching and assessing Communication.

One aspect of 'professionalism' about which we gathered evidence was the extent to which staff in the case-study college, and in the associated Regional group, were engaged in monitoring and reviewing their assessment practices and decisions. The staff in the case-study college tried to meet together as often as possible, though at the beginning of the academic session in which the research took place there were some time-tabling problems. There was also a great deal of informal discussion reported to us by the interviewees and the Regional group met regularly. If it were not for the existence of these arrangements, the research would have had access to much less information within the case-study college and in the other colleges. The member of staff from the case-study college who liaised with the Regional group saw this development as an important way forward and as a valuable source of support:

> I think we've still got a long way to go. I think we're progressing, but it's mainly in the area of internal assessment and moderation.
>
> It's taken us three years to really come to some sort of understanding about what the outcomes are asking for ... I think that's why so many groups have sprung up, because people were floundering in the early days, they found it so complex. They were really quite threatened by the whole situation, they didn't feel they could cope on their own.

Other lecturers, when asked if they thought that these group meetings had helped, welcomed them as a way of dealing with uncertainties thrown up by the descriptor and commented:

> I suppose I have only come to having some idea of what's the difference between [levels] 2 and 3 from discussion with other people in my department.
>
> Lecturers teaching this module meet together once a week to discuss problems, standards, moderation and this helps to ensure some degree of agreement ... However, greater guidance is necessary. One way to do this is to encourage time-tabled Communication meetings for staff; more exemplar material; internal moderation group to monitor assessment supported by regional moderation group.

Another source of reassurance for some was the use of 'double marking' or reviewing one another's assessments:

> We have already recommended that for level 4 certainly every student's work should be examined by more than one person.
>
> ... within the college, because we tend to double up when we're doing final assessments, especially in the oral area, staff feel a bit happier that they're having a second opinion.

Similarly many of the summative assessment decisions were based on the work that students had in their folders, and staff in this college reported reviewing such work with colleagues before reaching a final decision and, if necessary, asking that the student should submit further work for confirmation of their decision.

The care which staff in this college, and in the colleges associated with it in the Regional group, have taken to harmonise their understandings of the requirements of the descriptor may go a long way towards explaining the degree of comparability we found in their assessment decisions. Doubts remain, however, about assessment at the more detailed level of individual criteria within the sub-skills.

General Ability versus Particularised Skills

The National Certificate set out to produce precise statements about what individual students could and could not do. In the Communication descriptor, however, student achievements are reported at the learning outcome level while assessment is expected to take place at the sub-skill level, and no guidance is given as to how the sub-skill scores are to be aggregated into a level for the learning outcome. The suspicion arises that the 'collective subjectivity' which has been arrived at may be reflecting the lecturers' internal notions of some sort of general communication ability rather than the students' actual achievements on particular individualised skills.

In the course of the interviews, staff were asked about the way achievements were aggregated into overall learning outcome awards and whether it was possible that some form of 'trade-off' could occur so that achievement on one skill could offset lack of achievement on another. There seemed to be agreement that this was possible but was not a great danger as far as trade-off between the four modes or learning outcomes was concerned:

> I think this is less likely with modules than it was in Communication courses if only because you are being focused on, if you want, individual boxes and ticks — which probably means you have to actually think what you're ticking. Yes, it still exists, but I would say not terribly much.
>
> That does not happen because in Communication with the new system of recording we can give them a totally separate mark for each of the four modes and that is something we have welcomed.

However, if trade-off between modes was thought unlikely, there was a feeling that trade-off within the modes could, and did, happen, perhaps because one sub-skill within a learning outcome was considered to be of vital importance; because of a feeling that these skills were generalisable; because it was practically difficult to maintain an even balance; or even because there was a tendency to assess on the basis of the impression given by the student:

> There probably would be a tendency to carry ... that's an awkward one, actually. I was about to say that folk who could write the one could probably write the other — which probably answers your question.
>
> I think perhaps because there are so many sub-skills, I'm quite conscious that for some of them I do quite a lot of assessing and it's quite difficult to get the balance to make sure that I'm assessing them all evenly.
>
> When you put the tick in the box you think "What do I know about this student? I remember what he did, or she did, last week, oh yes, that's fine, I'll put a tick in the box."

This last quotation suggests an 'impression marking' approach which is reminiscent of the method used by the researcher in assessing the Listening test at the learning outcome level. There the learning outcome and performance criteria at the four levels were treated as broad descriptions of typical performance rather than strict criteria. The results showed a high correlation with those obtained by the college and suggest this may be the normal method of approach for some Communication staff.

Comments made by some lecturers seem to confirm this. One noted that the learning outcomes were not such a pressing concern now as they had once been:

> I think to start with, the outcomes were very much at the forefront of our minds ... I'm not so worried now about the outcomes as I was to begin with. I think it's better for the students and better for their education, the interest of the topic rather than the outcomes.

The same lecturer went on to comment that this was appropriate to the way she saw Communication as a subject:

> If I'm paying attention to each sub-skill I think it defeats the purpose of the module in that I think in all these modules there should be more general outcomes rather than just specific task-related outcomes ... I think that Communication ought to be looked at in that way, I don't think it should be fragmented.

Another lecturer expressed a similar view:

> I think it's the process which is the important thing in the Communication module, not the final summative assessment.

These lecturers clearly have their own view about the nature of their subject. If it is not entirely in accord with the assessment philosophy of the National Certificate then it also has to be said that neither is the Communication module descriptor which, as it stands, is a compromise between the breadth of the subject and the assessment requirements of the National Certificate.

CONCLUSIONS

It might be suggested that the Communication module descriptor could be modified to suit the needs of the National Certificate better than it does. Perhaps reporting could take place at the sub-skill level rather than the learning outcome level. Domain definitions could be tightened. Performance criteria could be stated more clearly and ambiguities removed. Perhaps some more prescribed form of assessment could be introduced or more advice given on what should be included in the students' assessment folios(and these possibilities were both suggested by lecturers in the course of interviews). However, experience with other, more 'tightly defined' modules does not lead us to think that this would remove all problems. Furthermore, we have noted elsewhere (Black, Hall *and* Yates, 1988) that tightly defined module descriptors have problems all of their own. Nor would a move in this direction be likely to go down well with teaching staff.

An alternative might be to simplify the tangle of sub-skills and performance criteria and accept that the level of particularity that they represent at present is impractical. Some guidance on how they should be aggregated into learning outcome awards would also be an advantage.

6 Electronics:observations on a practical module

THE ELECTRONICS MODULE

Module 64306, Basic Soldering and Diagnostic Techniques on Electronic Circuits, is classified as a specialist module within the electronics group of the National Catalogue. It is described as being a module 'which enables the student to acquire the basic skills necessary to carry out development, maintenance and tests in electronic circuits'.

The module was chosen partly to ensure that a practical module was included amongst our case-studies and also because we wanted to cover a wide range of recommended assessment procedures. Within 64306 two main types of assessment instrument are recommended: an observation checklist and examination of the finished artefacts. The latter include prepared drawings, assembled circuits, and finished articles. The product of practical work and observations made by checklist are frequently used instruments of assessment in the National Certificate.

INITIAL HYPOTHESES

The module descriptor suggests domains to be assessed which are manageable in size, competences to be assessed which are clearly defined, and content which is clearly stated. It was, therefore, a fairly 'tightly defined' module. There were some technical flaws relating, for example, to the use of loose adverbs in the performance criteria, but overall it was our expectation that staff would find the descriptor supportive of good assessment practice.

QUESTIONS FOR THIS CASE-STUDY

The feature of this case-study which makes it different from the others is that assessment was based on the ongoing work of the student with no recourse to formal 'end of module' tests. Such assessments are notoriously difficult and expensive to monitor but they offered us the opportunity to ask questions, which were different from those addressed by the earlier studies.

In particular, we asked about what goes on in the classroom. Such questions are of central importance in understanding the process of assessment, but are seldom addressed. We set out to find out what it was like to be a student on the 'receiving end' of the assessment process used by the department in this case study and to determine what incidents raised questions about the National Certificate assessment model. However, we also asked questions about staff experience of the module descriptor as we had in the other studies. In particular, we were interested in whether lecturers' interpretations of the advice offered in the module descriptor resulted in sound assessments. Where there was doubt about the adequacy of the process, what appeared to be the reasons behind it?

THE RESEARCH DESIGN

Research was conducted in two colleges. There are some similarities between the data gathered in this study and those for the other case-studies. Each member of staff who taught this module was interviewed using the common schedule and

completed a questionnaire. The interviews and questionnaire were also given to the staff of another college which worked with us, and data were collected by questionnaire from a national sample of staff teaching the module throughout Scotland.

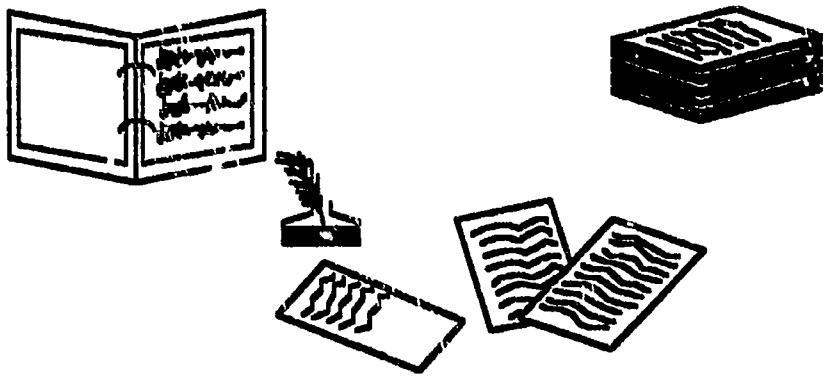
Of central importance to the design of this study, however, was participant observation. For this, a researcher was present as a student throughout the module in one of the colleges. He carefully observed and recorded what happened, how decisions were made, and the ease or difficulty encountered by staff in carrying out assessment. This approach offered a unique opportunity to see at first hand how this module was being assessed.

Negotiating a participant observer role proved to be a delicate task. From the outset, senior staff in the college were told that the primary purpose of researching the module was to gain an insider's experience of assessment and to enable comment to be made on how the instruments and processes of assessment were implemented. At a later stage, it was necessary for the researcher to reassure the lecturer concerned that the main focus of the enquiry would be on the potential of the National Certificate to produce reliable and valid assessments, and not on his particular strengths or weaknesses as an assessor.

At two points in the teaching of the module, the lecturer and the researcher carried out parallel assessments of student performance. Students were asked to complete self-assessment questionnaires on three occasions. But the most important set of data was the researcher's observations over the 40 hours when the module was being delivered.

Data from participant observation are reported in a different way from those in the rest of this report. Because the researcher was actively engaged in the work of the class, the emphasis in this case-study is not on quantitative results of assessment, but on the process of assessment. The most important findings arise from the researcher's awareness of critical incidents reflecting the quality of the process. Accordingly, what we offer here are extracts from a diary of experiences during the course of the module which illuminate some of the issues raised by this type of assessment. The extracts are not from every week but only for those which offer insights on the research questions.

The Research Diary



The Research Diary

Week One

All the students who had enrolled for the module were given a copy of the learning outcomes and a brief statement that outcomes would be assessed by class exercises and practical assignments. What was missing was any indication of what had to be done to demonstrate that the outcomes had been achieved. We needed access to the specific performance criteria for each learning outcome.

Comment

To get maximum benefit from criterion referenced assessment, all concerned should know the grounds on which decisions will be made. It was helpful to be given the learning outcomes, but this was only part of the message. How often must I show competence in a skill to demonstrate mastery?

Assessment started in the first week, after a very brief introduction to basic soldering (demonstration of techniques, introduction to the tools, provision of materials and instruction to complete an exercise on Learning Outcome 1). During the exercise, lecturer moved around the class of 12, inspecting tag boards and each student's soldering. For each one he demonstrated how to use a multi-meter and gave help of different kinds. He seemed to take account of students' occupations and background. Perhaps he had different expectations for each of us.

Comment

An element of formative assessment was involved, but it eventually built into a summative form. It gave me the opportunity to seek help, but if I were to seek it too often, would I be labelled as a 'non-master'?

The Research Diary

Week Two

The students in the class all knew they were being assessed. They had been told what were the learning outcomes for the module. But did they feel that they were meeting new challenges? Did they feel that they already had the skills they were about to be taught? And as the module progressed, would their self-assessment agree with the lecturer's views on their competence? I set out to track some of these issues systematically. The team devised a self-assessment questionnaire and it was applied this week. It would be applied again in weeks 10 and 15, (week 15 was near the end of module and at this point the number in the class had dropped from 14 to 10). It asked my student colleagues to rate their competence on each of the sub-skills embodied in thelearning outcomes.

The table attached gives the results not only for this week, but also for the other two applications. Would their self-assessment profile change as the course progressed?

Student Self-Assessment Grid

Learning outcomes	Performance Criteria	Week in which grid was administered: Week 2	Week 10	Week 15
1. Can you demonstrate basic soldering and desoldering techniques and consistently:	(a) prepare the material adequately	14Y	11Y	10Y
	(b) solder/desolder neatly and accurately	14Y	11Y	10Y
	(c) avoid damage to components or print strip	13Y 1N	11Y	10Y
2. Can you prepare the layout, assemble and test a printed circuit and stripboard assembly from a simple circuit diagram and show that:	(a) the prepared drawing is neat and accurate;	2Y 12N	11Y	10Y
	(b) the design grids are used correctly;	1Y 13N	11Y	10Y
	(c) the layout is designed for maximum stripboard economy;	14-N	10Y 1N	10Y
	(d) the P.C.B. artwork, etching, drilling is comparable with an agreed exemplar;	1Y 13N	11Y	10Y
	(e) the circuit [illegible]ons as required	1Y 15N	8Y 3N	10Y
3. Can you fabricate wire wrap connections and assemble ribbon cable connectors and:	(a) use the appropriate tools correctly and safely	14Y	11Y	10Y
	(b) produce good quality connections	4Y 10N	11Y	10Y
4. Can you select and use appropriate instruments to measure currents and voltages in pasive and active circuits and in a series of practical exercises:	(a) select the correct instrument and scales;	10Y 4N	9Y 2N	10Y
	(b) adjust the instrument correctly;	11Y 3N	9Y 2N	10Y
	(c) connect the instrument correctly to the circuit;	9Y 3N	9Y 2N	10Y
	(d) interpret scale readings correctly	10Y 4N	9Y 2N	10Y
5. Can you test simple combinational logic circuits using logic probes clips and pulsers and in a series of tests:	(a) use logic probes correctly;	14N	4Y 7N	10Y
	(b) use logic clips correctly;	14N	1Y 10N	10Y
	(c) use logic pulsers correctly	14N	1Y 10N	10Y
	(d) locate faults using probes and/or clips and pulsers	14N	1Y 10N	10Y
6. Can you show that you are able to use the oscilloscope in a circuit and in a series of tests:	(a) use the controls correctly;	14N	2Y 9N	10Y
	(b) connect the instruments correctly to the circuit;	14N	2Y 9N	10Y
	(c) take accurate measurements	14N	2Y 9N	9Y 1N
7. Can you show that you are able to use waveform generators in test circuits and in a series of tests:	(a) use the controls correctly;	14N	2Y 9N	9Y 1N
	(b) connect the instrument correctly and monitor the apporopriate display on the oscilloscope.	14N	1Y 10N	9Y 1N

Y = Yes, I think I have achieved the particular learning outcome

N = No, I do not think I have achieved the outcome

The Research Diary

Week Three

Several students encountered problems with circuits which were not functioning as anticipated. On investigation it appeared that heat damage to certain components eg transistors had adversely affected their performance.

Comment

False mastery classifications might be made as a result of malfunctioning circuits which had previously damaged components. The issue was of concern to staff and one lecturer drew my attention to it in an interview:

> "Quite often it's very difficult to get the practical devices to work, and if you don't get them to work after a bit of fault finding you've got a serious problem. There's so many things that may have gone wrong, and the components may be damaged. So you've got to say well, the guy's done everything, and he's built it, and he's had a go at testing it. Although it doesn't actually work, do we fail him or do we pass him?"

The Research Diary

Week Four

Learning Outcome 1 requires that the student should demonstrate soldering and desoldering techniques on listed items, the performance criteria being that the student consistently prepares the material adequately, solders/desolders neatly and accurately, and avoids damage to components or print strip. For part of the assessment students were required to complete soldering operations on a tag board to produce two artefacts, one comprising resistors, the other resistors, capacitors, and transistors. Assessments were made both by the lecturer and by the researcher.

Judgements were made by about neatness and accuracy of soldering, and about any damage to components and, where appropriate, tests were undertaken to ensure that the circuit in the second of the sub-tasks functioned as intended. In all cases lecturer and researcher came to the same assessment decision. Two examples of the comments made are given below.

Student	Comments of Lecturer	Comments of Researcher
A	Very neat. Solder joints are pretty good. All connections clear from evidence of burning. Sharper on corners.	Relatively neat lay-out. Soldering appeared clean - no evidence of burning of wire. Slightly long-wires protruding through.
B	Soldering is very good. Some shrink back on the PVC with the heat which was applied.	Clean soldering operations. Some very slight evidence of scorch marks to wiring in one place.

Comment

The examination of artefacts has become the normal means of assessing skills in this module. This means that decisions are based on retrospective assessments, eg evidence of burning marks, and summative decisions about the adequacy of soldering/desoldering skills. These are detached from any diagnostic assessment on the actual processing skills as they are being performed.

The Research Diary

Week Five

I took the opportunity of a slight lull in the pace of work to move around the class examining what had been produced by my fellow students. By now class members were well spread out in terms of what work each had completed and was now doing. Some were well advanced beyond my stage and had been able to complete work to a very high specification level. One of the class told me he had completed certain tasks and produced finished artefacts outwith the class contact time at his workplace.

Comment

The lecturer expected students to do "homework" to cover the work in the time allocated. A problem which arises when work which is to be assessed is completed outwith class time, is the assurance that the work is genuinely the student's own.

This week I approached the lecturer about a practical difficulty I was experiencing in the soldering operations. I was shown how I could attain some degree of stability by means of balancing pliers on the side of the board.

Comment

This advice or "trade tip" was a clear example of effective diagnostic assessment by the lecturer. It had a crucial bearing on the quality of the artefacts I subsequently produced. I wondered to what extent diagnostic assessment of this kind might become synonymous with summative assessment, and any distinction between these two assessment purposes become blurred.

The Research Diary

Week Seven

The attainment of Learning Outcome 3 involved the lecturer in taking small groups of students to demonstrate and explain the operation of a ribbon cable connector machine. After outlining the stages to complete ribbon cable connection we were asked to work in pairs to produce connectors and to test for continuity on a 12 pin plug connector. There did not appear to be any formal assessment of this operation beyond the opportunity to use the machine and to test for ourselves the continuity of the connectors. Subsequent discussion with the lecturer revealed that he considered this learning outcome to be anachronistic and unnecessary in terms of current industrial practice, and saw no good reason for spending a lot of time on it.

Comment

Lecturers should exercise their discretion in the time they give to different learning outcomes. However, it must be debatable as to whether the most appropriate strategy is unilateral dismissal. If learning outcomes appear to be redundant to current needs, it would be highly desirable to record this with SCOTVEC as it is only through this channel that module descriptors will be improved.

The Research Diary

Week Eight

As part of the assessment for learning outcome 2 the class were asked to produce a logic probe having specific physical dimensions. All the logic probes functioned as required, and the criterion of achieving maximum stripboard economy was defined as the ability of the probe to be inserted into a plastic tube of a pre-determined size.

Comment

This artefact was the second of four PCBs produced by the students, and although it provided a part of the summative assessment for learning outcome 2 opportunity for improvement was given in two further practical exercises. This is interesting because there is no indication given in the descriptor of how many prepared drawings or circuits it is necessary for the student to produce in order to achieve this outcome. The module descriptor could be interpreted by a lecturer to allow just one drawing/ circuit as the instrument, and does not indicate the degree of complexity of circuitry which is to provide the basis for the design and construction of the PCB.

The Research Diary

Week Nine

We are more than half way through the module, and in keeping with most of the class I am still working on learning outcome 2. We have been given boards on which components and circuit lay-out are to be designed and produced to attain the criterion of "maximum stripboard economy". One student has produced a highly original and striking double image circuit which reproduces the face of a cat in flashing lights on the LED chaser circuit. A considerable amount of complex design work has gone into the production of this artefact. The module descriptor does not specify the level of complexity of circuitry or design lay-out for this outcome to be achieved, though all students in our group are working to the same framework of a circuit diagram and size of PCB.

Comment

A "minimum level competence" requirement has been given us in the parameter of size and circuit diagram, but some students have exceeded this, perhaps because of their greater background experience. In keeping with the spirit of student-centred learning and the need to retain the interest and involvement of students in the voluntary evening class context, students are being encouraged to perform to their own level of competence and potential for progress. I wonder whether under these conditions summative assessment carries with it a strong ipsative dimension, and whether some students might feel themselves being discriminated against by being assessed in terms of different interpretations of the same performance criteria.

The Research Diary

Week Ten

This week the students completed the second self-assessment grid. The results have been recorded in the diary for week 2 to make comparison easier. The data indicate that the students thought that they were gaining new competencies as the module proceeded. By this time in the delivery of the module there is a very wide spread in competence within the module and this has resulted in individuals being at various stages of achieving all the learning outcomes. The performance criteria for each outcome comprise several tasks, and staff have comprehensive observation checklists which extend over several pages to assess outcomes. Some students who are well advanced in the module have been given extension work in the form of more sophisticated lay-outs or in repeating their own circuits as demonstration models. Other students, on the other hand, are finding it quite difficult just to keep up with the work schedule.

Comment

Attention to the particular needs of and challenges for the individual student places very stringent requirements on the member of staff delivering a module. They need to be well organised and to have well maintained and up-to-date records of what students in a group have achieved to enable them to make helpful diagnostic as well as necessary summative assessment. Support might be given in the shape of a well-designed and easy to handle record systems for student attainment.

The Research Diary

Week Twelve

During most of the course we have not been pressurised to complete tasks within a specific period of time. However, as we approach the end of the sixteen week block, there is some urgency to complete particular learning outcomes so that, at least in theory, there might be time for any necessary remediation and re-assessment. I have been asked to complete my LED chaser circuit within the near future.

Comment

The introduction of time limits within which tasks might be expected to be completed could create scope for differentiation in the assessment of students. Within any commercial manufacturing context it is normal practice for time constraints to operate in the performance of tasks. Consequently it would be quite consistent with current industrial practice for time limits to be introduced into the assessment of practical tasks in modules. However for this to be acceptable it would have to be identified as part of the performance criteria within the descriptor.

The Research Diary

Week Fifteen

This week the students completed the final self-assessment grid and the results are again recorded in week 2 of the diary. In addition they answered a small number of more general questions relating to their perception of the value of specific performance criteria in relation to actual or intended work, their awareness of when and how they were being assessed, the information conveyed to them as they were completing the module, and a general question about some aspects of their experience of being assessed in National Certificate modules.

The data are reported below.
Clearly most students are not aware of when they are being assessed, what instrument is being used, or even with what results, in their progress towards overtaking learning outcomes. However, a group of questions relating to the students' confidence that they had gained authentic mastery of the performance criteria indicated that this confidence was very high. Students also related assessment in National Certificate modules as being easier, more informative, more helpful to future progress, and fairer than other forms of assessment which they had experienced.

Results of Student Questionnaire

1. Do you think that the knowledge and skills you are gaining by doing this module are helpful to you in your work?

   YES [8]
   NO [2]

2. Are you aware when you are being assessed in this module?

   YES [2]
   NO [8]

3. Are you aware how you are being assessed in this module (eg. that you are being assessed using an observation checklist)?

   YES [1]
   NO [9]

4. Are you told that you have achieved a learning outcome or will be required to repeat an assessment in the module?

   YES [5]
   NO [5]

5. Were you confident that when you achieved a particular learning outcome in this module you really had acquired the skill or competence indicatedby the corresponding performance criteria?

   YES [10]
   NO [·]

6. Are you confident that if you have achieved a particular learning outcome in this module, you have acquired the skill or competence indicated by the corresponding performance criteria?

   YES [10]
   NO [·]

7. Compared to other forms of assessment which you have experienced do you find that assessment in National Certificate modules is (please tick):

Easier	[8]	or	More difficult	[2]
Less informative	[3]	or	More informative	[7]
More helpful to my future progress	[7]	or	Less helpful to my future progress	[3]
Less fair than other forms of assessment	[0]	or	Fairer than other forms of assessment	[10]

The Research Diary

Week Sixteen

We are in the last week, and most of us who have lasted the course, produced the required artefacts, and completed all the practical exercises and tasks, now know that we will be very likely to achieve all the learning outcomes of this module. However, as a check our lecturer consulted his own record sheet to confirm student by student which tasks had been completed satisfactorily, and which required to be finished. All students present were asked to hand in any worksheets still outstanding.

Comment

Remediation throughout the module has been on an informal basis. Many students will be carrying on to complete a more advanced module. This would make it simple to conduct any necessary remediation and re-assessment of the learning outcomes of this module. However, for those students not progressing to a further module there was no obvious opportunity for remediation and re-assessment within the forty hours delivery time for this module, though a formal procedure was available through the department.

All students should be made aware of opportunities for remediation and re-assessment given the time constraints involved in modules like 64306, in which students are under considerable pressure to complete a tight schedule of assessment exercises and tasks.

DISCUSSION

In this case-study the adoption of a participant observer role afforded the researcher an opportunity to focus on the adequacy of the process of assessment. These observation data, combined with those from our interviews and questionnaires, allow us to comment on each of the matters considered in the other studies.

The Quality of the Instruments: Precursor Variables

Adequacy of the domain definition

In our view, the domains set out for assessment in this module were clear and manageable. Although staff comment seemed to confirm our views, there were a few negative comments about details of particular outcomes: learning outcome 5 was seen as lacking clarity and learning outcome 2 was felt to be too broad and hence could only be handled in a shallow (and unsatisfactory) way.

> I think (the module as a whole) compares favourably with other module descriptors.

> I would say the whole content and context of that module is as right as it can be seen to be.

Week 7 of the research diary however, provides a slightly different insight into the question of which domains to assess. A potentially important variable in determining the quality of

assessments may be the attitude which staff have to the learning outcomes they are asked to assess. In week 7 it was noted that because the lecturer was unconvinced of the appropriateness of outcome 3 as described in the descriptor, he chose to treat its assessment very lightly.

Clarity of what is meant by mastery
Questions about the clarity and appropriateness of the conceptualisation of mastery within the module descriptor elicited more scepticism on the part of staff:

> My interpretation (of master) is that somebody is totally competent and confident across all the areas. I don't think that's what we're doing...

> The level of achievement must really be supplied by the lecturer, which depends on the lecturer's own experience in this area.

> I always query 'adequately', 'neatly' and 'accurately' ... I think we all have our own standards of neatness and accuracy.

These data suggest that staff had encountered problems in making decisions about mastery when the performance criteria are qualified by such adverbs as 'adequately' and 'accurately'. At the same time, the parallel assessments carried out in weeks 4 and 8 (and reported in week 4 of the diary) indicated a high level of agreement. It may be that although staff would feel more secure if the performance criteria were tighter, the apparently sound domain definitions in the descriptor make the recognition of acceptable performance easier than the lecturers realised.

Adequacy of the instrument specification
The help offered on the construction of instruments is minimal. Observation checklists, the examination of finished articles and the preparation of drawings are all prescribed, but without further elaboration. However, the fact that instrument specifications are dealt with in juxtaposition with the performance criteria helps to clarify the instruments' format. Moreover, comments from staff suggested that they did not clearly distinguish between the two. Comments made about the assessment instruments suggested that there was dissatisfaction with them.

Only one member of staff teaching this module expressed confidence that the instrument specifications were adequate, and even his views were qualified:

> The detail is specific enough, but it would be much more difficult for somebody who didn't have the experience and practical background. I have seen other colleges' instruments of assessment. There are quite wide variations.

While there was little support for the quality of the advice provided, few suggestions were offered on how it could be improved. Indeed there were contrary views on the feasibility of improvements, with one lecturer claiming that the nature of the instruments specified made support difficult, and another suggesting that the solution lay in the provision of examples of good practice.

I don't think you are ever going to get general agreement. But you would get agreement if an exemplar was produced for a particular thing.

When it is an observation checklist or a professional judgment on a finished article, then I'm not sure how much guidance can be given.

The Quality of the Instruments: Operational Variables

Adequacy of the components of the instruments

Assessment 'items' in this context comprised the individual observations and judgments about artefacts which accumulated to inform the decision made by the lecturer about his students' attainments. It is just as important to ask whether each of these is an adequate assessment of the domain being assessed as it is in the context of 'pencil and paper' tests, but that is not an easy exercise and the technology associated with it is ill-developed. Also, because these essentially ephemeral interactions are so inaccessible, there has been little attempt in the past to monitor their quality.

Our participant observation design offered the opportunity to examine such 'items' in more detail. In the main, it appeared that there was a high fidelity between the individual observations and the skills required in the learning outcomes. However, as was noted in week 4, there were some instances where outcomes which should have been assessed through the observation of process seemed to be evaluated through examination of the end product. In several instances extraneous factors were introduced into the system. In week 5 students may have received help from others when work was completed outside the college. In week 12, time limits were introduced for the first time and their status in relation to the performance criteria was unclear.

The main cause for concern about the quality of the individual components of the assessments, however, arose from the demands made on the lecturer. Checklist assessment is not easy and was particularly difficult to manage in week 10, in the context of individualised learning. We have no evidence to suggest that the individual lecturer was other than effective in his application of the approach, but we could discern no parallel to the systematic review of practice in our Communication case study. In a context where quality control is so difficult to implement, this may have been a major weakness.

Adequacy of the instruments

Learning outcome 1 requires students to fit and replace components on a pre-constructed board. However, as staff pointed out to us, the number of components, the size of the board and whether it is stripboard or printed circuit, could have a significant influence on the difficulty which students might have in completing soldering operations. Decisions on these matters are left to the professional judgment of the teachers. Moreover, the number of operations required to demonstrate a consistency in performance is also left to their discretion.

The number of components used, and the number of instances in which a student is asked to display competence, are the parallel concerns to the 'test length' problem which was discussed in the Mathematics and Stock Control case-studies. The associated

question of whether the tasks included adequately sample the domain being assessed also has parallels. We found that the number of observations used to arrive at assessment decisions seemed to vary. We say 'seemed' because it is almost impossible to decide when a lecturer is making such an assessment. For example, in week 5 the researcher suspected that the formative help given to him in soldering may also have constituted the major summative assessment of his competence in that skill. Had soldering been a weakness it would have shown up in later work and alerted the lecturer to the problem.

Decisions issues

Despite these findings and the critical comments made on some aspects of the precursor variables, our findings suggest that the decisions on mastery made by the lecturer seemed to be sound. There were a few instances, however, where extraneous factors may have influenced thinking. On one occasion, for example, the lecturer appeared to be influenced by the neatness of a particular piece of work when this was not explicitly mentioned in the performance criteria.

It was not only our observations which supported this conclusion. The parallel assessments carried out, and the lecturer's assessments of student attainment and their own self-assessments by the end of the course, showed substantial agreement. Staff were not sure, however, that this could be attributed to the quality of the module descriptor. In their estimation it was at least partly explained by their 'professional judgment':

> It comes back to your own personal experience and knowledge. Therefore if these are areas in which you yourself feel confident and familiar, then you do feel you can make your own judgments.

The Quality of the Process of Assessment

The assessments observed during the research were based on actual student performance, and all students were given an equal opportunity to display the minimum levels of competence on which mastery decisions were made. We were not always sure how many 'items' or observations were used to arrive at decisions on mastery, but it seemed that the lecturer in our case-study paid due attention to the fundamentals of 'professionalism' when making his assessments.

In a context where process and observational approaches are the basis of assessment, we cannot distinguish between the way in which the 'instruments' are used to arrive at mastery decisions and the 'process' of making decisions. There were, however, some aspects of the process of assessment which brought into question the full extent to which it fitted the philosophy of the Action Plan. Some of these have implications for the quality of the decisions, while others have significance for the relationship between teaching, learning and assessment.

In week 1, students were not given the performance criteria on which they were being judged and were not always aware of when and how they were being assessed. In week 15, there was an absence of clear and unambiguous feedback to students on which learning outcomes they had achieved as they completed their work. Each of these findings runs counter to the basic premise in the

Action Plan philosophy that students should share with their teachers a clear understanding of what has to be achieved in order to succeed. It may well be that knowledge about progress towards clearly understood aims has a motivating influence. If staff do not value sharing criteria in this way it could have implicatio. for student attainment.

Action Plan philosophy, especially in relation to student-centred learning, was understood and implemented by the lecturer. Tasks were directed towards the particular needs, interests and abilities of individual students. The typical form of pedagogy was a one-to-one relationship between the lecturer and the student. But it is in the relationship of the assessment proces. to this style of teaching and learning that our second set of findings have relevance.

We noted in week 16 that it was not always clear to students that there was an opportunity for formal assessment of progress and 'remedial' support in cases where problems were encountered. This is hardly surprising, as the students themselves were unaware when they were being assessed and what their progress, as seen by the lecturer, was. Furthermore, from week 1 onwards the participant observer noticed a tension between the formative and summative purposes. Would 'owning up' to learning difficulties count against the student in the summative context? A clearer understanding between students and teachers about diagnosis and remediation might have paid dividends, if not in improving the quality of summative assessment, then certainly in enhancing the learning experience.

CONCLUSIONS

In this case-study, questions were posed on two broad themes. What it was like to be a student on the 'receiving end' of the assessment process? And could we identify incidents which would help us to understand what determined 'quality' in this context? We also wanted to learn about staff experience with this particular module descriptor.

First, it is worth noting that the student questionnaire of week 15 indicated a positive student attitude to the way assessment was conducted, but there was evidence that not all aspects of what might be termed 'best practice' in National Certificate assessment had been achieved. In particular, students lacked information about what they had to do to succeed and about their progress.

Because all the students were successful, we were not able to analyse the views of those who failed to master outcomes, nor to identify with the lecturer the reasons why such decisions had been made. This has limited the insights gained from the student perspective. However, the positive student attitude encountered here reinforces the findings from the earlier stages of the project (Black, Hall *and* Yates, 1988).

Our findings about the way in which the module descriptor supported assessment were generally positive. We believed that the domains identified in this module were relatively clear and manageable and staff comment confirmed this. Problems had been encountered, however, in making decisions about mastery when the performance criteria were qualified by loose adverbs, and there was little confidence that the instrument specifications were

adequate. Overall, this descriptor had fewer problems than those in our other case studies, and we suspect that much of the reason for this was the realistic scale and high specificity of the domains identified in the learning outcomes.

The essentially ephemeral interactions which characterise assessments of the kind required in this module are notoriously difficult to evaluate although our participant observation approach did provide some insights. There appeared to be a high fidelity between the individual observations and the skills required in the learning outcomes, and the decisions on mastery made by the lecturer seemed sound. However, staff claimed that this was as much due to 'professional judgment' as it was to the quality of the module descriptor. In circumstances where 'quality control' is so difficult to implement, it would have been encouraging to have found evidence of more attempts to review the assessment procedures.

7 Conclusions

In this chapter we will attempt to pull together the various threads of our enquiry and discuss the broader implications of our case study findings. We hope that some of what we say will strike Further Education staff as familiar and help to sharpen insights into the problems of assessment in the National Certificate.

THE RESEARCH QUESTIONS

In the introduction to this report we established two fundamental questions for our study. The first was to explore the soundness of assessments made in the case-study departments. The second was to consider the various reasons there might be for any differences in quality. In each case-study these basic questions were reinterpreted to embrace the particular features of the module or the subject area under consideration.

HOW SOUND WERE THE ASSESSMENTS?

In Chapter 1, the four features identified as characterising sound assessments in the National Certificate were that they should

- be carried out in a professional and unbiased way;
- adequately reflect the specific learning outcome they purport to measure;
- allow users to arrive at clear and accurate decisions about students' attainments;
- be reliable indicators to end-users of the particular message that the National Certificate wishes to deliver.

Staff were generally professional in administering assessments

The findings suggest these goals were attained with varying degrees of success. In all of the case-studies, staff were carrying out their responsibilities both professionally and in ways conducive to fair assessments. Of course there were problems. In the Electronics case-study the process could have been improved if the lecturer had made students aware of the performance criteria. In several instances, staff could have been more aware of the importance of reviewing their procedures systematically. But, overall, it did not appear that problems about quality could be explained by teacher-bias or a lack of professionalism.

Assessment *items* were valid in the sense of testing learning outcomes but there was some doubt about the validity of *instruments* because of the size of the domains

Whether the assessments adequately reflected the learning outcomes they purported to measure is a more difficult point. In Electronics and Communication it appeared that assessments were valid, but it was difficult to be certain. In Stock Control there was greater doubt about some of the instruments than others. In Mathematics, while individual items seemed to be sound, we had doubted the feasibility of making sound judgments when the domains being assessed were so broad and the tests being used were so brief. It seemed that while items were sampled adequately and appropriately from the domains in most instances, there was more doubt about whether alternative versions of instruments were doing the same job. This was probably because the domains being assessed were so large in size.

The question of whether the procedures used in the case study departments allowed users to arrive at clear and accurate decisions

The level of agreement between alternative assessment instruments was lower than anticipated in the 'hard' modules - Maths and Stock Control - but was higher than expected in Communication

about students' attainments was dealt with in some detail, but the outcome was somewhat ambiguous. There was a low level of agreement between the two forms of Mathematics test relating to A2, and although the situation relating to G2 was better, there was still considerable disagreement. The exchange of instruments between colleges in the Stock Control study yielded a much lower level of agreement than we had anticipated, and there was even less agreement between the two forms of instrument derived by the same staff in one of the colleges. In contrast to this we found a high level of agreement between alternative instruments in the Listening study and between different assessors' judgments in the Writing study and in Electronics. This would suggest that arriving at decisions which will stand up to external scrutiny is feasible.

Reliability would be enhanced if it were made clear whether 'master' means 'can do' or 'has done'

Sound assessment should, of course, be a reliable indicator to end-users of the particular message that the National Certificate wishes to deliver. Our research did not focus on this matter, but we can say that we are unconvinced that staff knew what it is that the end-users want to know. There is uncertainty even at the level of whether National Certificate assessments should indicate that a student 'has done' whatever is required of the learning outcome, or whether it implies a more demanding 'can do' message.

REASONS FOR DIFFERENCES IN QUALITY

Considerable emphasis has been placed on trying to understand why the assessments we studied varied in their technical quality. We will again use the Q-model identified in Figure 2.1 and used in all our case-studies, to draw our findings together.

Our approach does not claim to be fully comprehensive. Both the quality of curriculum analysis and variations in college, regional and national quality control mechanisms have a bearing on quality, but were outwith the remit of our study. Ours is only one of many possible approaches to unravelling the complex web of interactions between teacher, learner and 'system' in National Certificate assessment. It had the advantage of being grounded in existing knowledge about criterion-referenced assessment.

Precursor Variables

Staff in each of the case studies reported difficulties in clarifying the exact nature of the domains to be assessed

Our case-studies dealt with a range of subject areas, some of which were apparently more capable of precise definition than others. We also considered a range of assessment instruments, from multiple-choice exercises through to the observation of process skills. What is most noticeable is that there were problems with the domain definitions in each module, so it was not always clear what was to be assessed. These problems were most apparent in the Maths and Stock Control studies.

Why did we discover such frequent difficulties in working with the domain definitions in these modules? First, our research design may have highlighted them. Second, in some circumstances at least it is inherently difficult to formulate precise behavioural statements. And finally, the way in which modules have been constructed may have exacerbated the problem.

It has to be recognised that while it may be possible to define some domains to a high degree of specificity, this is not necessarily realistic

By claiming that some domain definitions were 'loose' we are assuming that greater clarity may be possible. It may not be possible, however, to achieve that clarity. Even in areas where it may be theoretically possible to produce exact specifications, these could become unwieldy and impractical. An example from our

case-studies would be the production of a printed circuit board in the Electronics module. To make absolutely clear and unambiguous what was expected of the students, statements about the type of circuit which was admissible, the size of the printed circuit board, the number of components to be used, the type of components, the complexity of the circuit, what materials and tools were to be available to the students, what could be considered an 'acceptable' layout, what quality of soldered joint was required, whether any scorching or discolouration would automatically disqualify the students (and if not, how much is allowable?), would have to be made and no doubt there are other factors which a subject expert could supply. This is only one learning outcome from a module with seven learning outcomes, and it will be clear that the resulting set of specifications would become very cumbersome. The fear is that no-one would read them.

Some domains are particularly difficult to define in advance

A second reason for not achieving clarity is that it may not be feasible to define some domains to this degree of specificity. This may be especially the case with learning outcomes which demand some form of 'open-ended' response from the students, and with the area of 'soft skills' or interpersonal behaviours. In these circumstances it may be possible to say what *type* of behaviour is required, and to recognise it, but not to specify in advance what it should consist of in other than very broad terms. Such learning outcomes were most apparent in the Communications modules. Examples would be the sub-skills within the Reading learning outcome which demands that the student 'evaluate the effectiveness of a communication', and within Writing which states that the student should 'employ forms of communication appropriate to purpose and audience'. In certain circumstances it may be possible to produce assessment instruments for these sub-skills in which acceptable student responses can be defined in advance clearly and unambiguously. However, it is usually difficult to define in advance the limits of what constitutes an 'evaluation' or what is an 'appropriate' form of communication.

There is a tension between the flexibility built into the National Certificate and the notion of having tightly defined domain definitions

Finally, there are certain policy decisions, taken at the very inception of the Action Plan, which seem to necessitate fairly broad definitions of the domains to be covered in modules. Once it was decided that National Certificate modules should be 'flexible' enough to cover a range of vocational areas, it followed that learning outcomes and content specifications in particular would have to be defined in fairly general terms, in order to leave room for the differing interpretations which separate vocational areas would require. Inevitably, this leads to a greater reliance on the 'professional judgment' of the teaching staff. This is especially so when we add the requirement that teaching and learning should be relevant to the students' vocational needs: a potential medical secretary will need a different Communication input from someone doing Media Studies. Unless there are to be different modules in these subject areas for different vocational groups, then the domain definitions within the modules will have to have room for manoeuvre. 'Room for manoeuvre' usually also means a certain amount of slackness of definition.

Operational Variables

When we move on to consider the factors which come into play when assessment instruments are actually used, it becomes more difficult to generalise across the case-studies. Problems tend to be

specific to the modules or types of assessment instrument in use. However, there are some issues which may be indicative of more general problems.

The looseness of the domain definitions was highlighted when the items were reviewed

Given the looseness of domain definitions, problems in matching individual assessment items to particular learning outcomes are hardly surprising. This was particularly apparent in the Maths study and especially the Stock Control study where we found substantial disagreement among the experts as to which learning outcome individual items were assessing. With individual checklist items and sub-skills in Communication, we also noted some disagreement about whether an observed behaviour belonged to one sub-skill or another.

In many cases assessments were based on a smaller number of items or observations than seemed to be appropriate

One major problem was that insufficient items and/or observations were used to guarantee that students really had mastered the domains assessed. This was partly because the domains tended to be so large that any adequate sample of items would result in over-assessment of the students, but also because some instrument specifications required as few as two items to assess a particular sub-domain. This is highlighted where multiple-choice or short-answer tests are used and must cast some doubt on the value of the resulting decisions. In Communication, the sampling took a less 'official' form, and on occasion certain sub-skills were deliberately under-emphasised for some student groups while others would receive greater attention. This is hardly surprising given the number and complexity of sub-skills in Communication and the lecturers' desire to concentrate on those of most relevance to their students. Indeed, it could be argued that this is evidence of good teaching practice. However, it again raises the question of whether a module taken in one context is 'the same' as that module in a different context.

The sound *teaching* practice of relating assessments to the needs of particular student groups can create problems for *assessment*

The nature of the advice given on aggregation of assessments within outcomes varied and in some cases the lack of advice created problems

Aggregation of assessments in different parts of domains was also a problem. In Communication, no guidance was given on how to aggregate the assessments for sub-skills into a learning outcome award and in other modules simply suggested the number of items to be used for each part of a learning outcome. The aggregation across learning outcomes in the Maths study compounds the problem but was, we suspect, an isolated aberration. This aggregation problem, combined with the way domains are sampled, means that students can achieve a learning outcome award without having achieved some component element. Worse things happened in traditional norm-referenced exams, but criterion-referenced assessment makes different claims: it is supposed to provide specific information about what a student can and cannot do.

Although the review of assessment instruments was observed at lecturer and college level there was, overall, a lack of systematic review

Finally, one feature of the data has greater implications for staff than for module descriptors: the variation in the extent to which the instruments and procedures were reviewed by staff. In some cases, notably the assessment of Talking in our Communication study, staff were supporting each other in evaluating their effectiveness. The Communication study also had an apparently effective system of support both within the college and at Regional level. And amongst the reasons the Maths study college had for working with us was their commitment to look systematically at the efficacy of their strategies. Despite this, there were many instances of assessments based on unreviewed procedures subsequently found to be lacking. Review may be

...nsuming initially, but it has the potential to improve the ...y of individual instruments, and to sharpen thinking more ...nerally. Overall quality could only benefit from it being taken more seriously.

Process Variables

There were problems with some of the assessments in our case-studies and many appeared to stem from the detail supplied by the descriptors, although it has to be underlined that the broad framework appeared to be sound in terms of assessment technology. Despite this, our impression is less negative than might be expected. The reasons for this become apparent when we turn to the process variables: that is, the components which relate to how staff actually carried out their assessments.

There are problems in this area. For example the way that attainments are aggregated has led to some form of 'trade-off' between learning outcomes, or between elements of a learning outcome. Despite this, the National Certificate demands and gets assessment based on attainment of learning outcomes rather than impressions of overall ability and, as such, was generally understood and welcomed.

Despite some evidence of 'trade-off' taking place, the notion of assessment against specific criteria was understood and accepted

There were also cases where it was not clear if the underlying basis on which mastery was to be determined was fully understood. This was noticeable in cases where cut-scores were used. For some staff the cut-score by itself was the performance criterion. Others appealed to industrial or commercial criteria for their decisions.

In some cases, 'mastery' was thought of only in terms of an isolated 'cut-score'. Some staff felt that their own commercial criteria were a necessary supplement to those provided

We found a substantial commitment amongst staff to make the system work. Most of those we worked with took a professional attitude to assessment and were aware of their responsibilities. Despite some inadequate guidance, most took great care over the assessments. Where there were doubts about what the module descriptors were asking of students, the staff tended to emphasise the preparation of the students for their vocational needs. There was little evidence of 'teaching to the test' or 'coaching' students through learning outcomes.

Staff were committed to the assessment procedures and were professional in meeting assessment requirements

Despite the inadequacies in the system, especially in the precursor variables, the conclusion we come to is that it is possible to produce consistency and comparability in assessment. Nowhere was this more evident than in the Communication case-study. This may be explained in part by the commitment to internal and inter-college monitoring and moderation. The 'looseness' of the Communication module descriptors may have acted as a spur to this. The Communication lecturers were aware that they were dealing with an ill-defined area and to have any chance of comparability between lecturers (and between colleges) they would have to consult one another. By doing this they could reach agreement about the meaning of the learning outcomes and performance criteria in their subject area. This takes a lot of time and effort, however, and not all the problems have been solved; but this area, which has the least precisely defined descriptor of any in our case-studies, and contains the greatest scope for subjective judgment, has produced the greatest consistency in decision-making.

The evidence would suggest that despite the problems we uncovered, it is possible to produce comparable assessments from the information available. However it is worth noting that the most notable example of high comparability occurred in the study in which we encountered the greatest amount of collaboration amongst staff

We shall always need the professional judgment of lecturers; it is

Descriptor guidance and sound professional judgment are necessary features of good assessment

unrealistic to expect that the information in the module descriptors alone will guarantee consistency and comparability. However, if the commitment exists, and if the resources and time are made available, the necessary professionalism of the teaching staff is available.

AN OVERVIEW

The problems identified in the case studies can be divided into two categories: *inevitable* and *avoidable*. They operate at three levels, 'policy', SCOTVEC and college:

We are aware that the possible reasons for variation in quality which we have offered are not the whole story, nor can we be certain about their relative importance. However, they operate at different levels and, at each level, some can be regarded as *inevitable* and others *avoidable*, depending often on decisions taken at a higher level.

At the highest level, decisions can properly be categorised as 'policy' issues, some stemming from the original Action Plan. These include the decisions that 'local flexibility' should be built into the National Certificate and that modules should be capable of being applied to different contexts, and to use observed rather than true scores (insofar as this does not appear to have been an option open to anyone operating at a lower level).

At the second level are issues relating to module descriptors formulated by SCOTVEC and roughly equivalent to our precursor variables in the Q-model. Included here are the very broad definitions and vast content of some domains, the poor definition of some learning outcomes, the rather weak conceptualisation of mastery in some module descriptors, and the inadequate guidelines on instrument specifications and decision making procedures in some descriptors.

Finally, at the college level are issues which arise when assessment is put into practice, and are roughly equivalent to our operational and process variables. They can be summed up as the need to rely on the professional judgment of lecturers, the problems which arise from inadequate domain sampling, the flexibility in the decision making procedures which encourages spurious borderline decisions, and difficulties caused by aggregation of scores (either within learning outcomes as in Communication, or between learning outcomes as in Mathematics). The problems which appear to emanate from each of these sources are summarised in Table 7.1.

From different points of view these problems appear as either *inevitable* or *avoidable*. For example, if we accept the policy decision that the National Certificate should strive for local flexibility and modules should be capable of being used in different vocational contexts, then it seems inevitable that domain definitions will have to be expressed in broad terms. This will lead to difficulties in providing detailed guidance, staff will have problems in sampling from these domains and everyone will rely on their professional judgment in making mastery decisions. These case-studies suggest 'professional judgment' is at its best when supported by active monitoring and reviewing. If it is not so supported, then inconsistency and incomparability of decision making are likely.

Conversely, if we wish to look just from an assessment perspective, the over-reliance on professional judgment and the difficulties in domain sampling can be overcome. But only if

Table 7.1 Problems in the Quality of Assessment

Policy Problems (National)	Module Descriptor Problems (SCOTVEC)	Operational and Process Problems (College)
• Local flexibility built into the National Certificate	• Broad domain definitions	• Aggregation of scores in order to arrive at mastery decisions
• Modules applied to different contexts	• Poorly defined learning outcomes	• Over-reliance on professional judgment of staff
• Decisions made using observed rather than true score	• Weak conceptualisation of mastery	• Problems arising from inadequate domain sampling
	• Paucity of guidance on instrument specifications and decision-making procedures	• Spurious borderline decisions

decisions are taken at a 'higher' level to tighten up the domain definitions, instrument specifications and other guidance within the module descriptors. And that may imply a decision at a still higher level, that modules should be less flexible and more context-specific.

IN CONCLUSION

We have identified three components of the system which, in our view, were responsible for the quality of assessments in the National Certificate. These are the institutions and procedures which have responsibility for policy at a national level; the process, largely with SCOTVEC, which determines, for example, the form and function of module descriptors; and the policies and practices within colleges, departments, classrooms and workplaces where the assessments actually take place.

Our case-studies have highlighted both successes and problems at each of these levels. At the level of national policy, support was evident for the broad directions taken by the National Certificate, but there were also questions about how to reconcile the tensions between the flexibility of modules and the implications of that for precision. For SCOTVEC, the fundamental structure of module descriptors seems sound, and in some case-studies proved to be the catalyst for assessments of high quality. In other instances, revision of the specific context of the descriptors seemed advisable, and we gather that such a programme is under way. At the college level, our case-studies pointed up the professionalism of the staff involved, but they also identified differences between institutions in the strategies they have for moderation and review.

The National Certificate is new and innovatory. It would have been strange indeed if in considering the quality of assessments we had found that all was perfect. If the account we have given in this report can be generalised to other cases, we hope that the insights we have gained may be of value to others who are involved in its implementation.

References

BERK, R A (Ed) (1980) *Criterion-Referenced Measurement: The State of the Art.* London: John Hopkins University Press.

BLACK, HD *and* DOCKRELL, WB *Criterion Referenced Assessment in the Classroom.* SCRE, Edinburgh.

BLACK, HD, HALL, JC *and* YATES, JB (1983) *Assessing Modules: Staff Perceptions of Assessment of the National Certificate.* Edinburgh: SCRE.

POPHAM, WJ (1978) *Criterion-Referenced Measurement.* Englewood Cliffs, New Jersey: Prentice-Hall.

ROID, GH *and* HALADYNA, TM (1982) *A Technology for Test-Item Writing.* London: Academic Press.

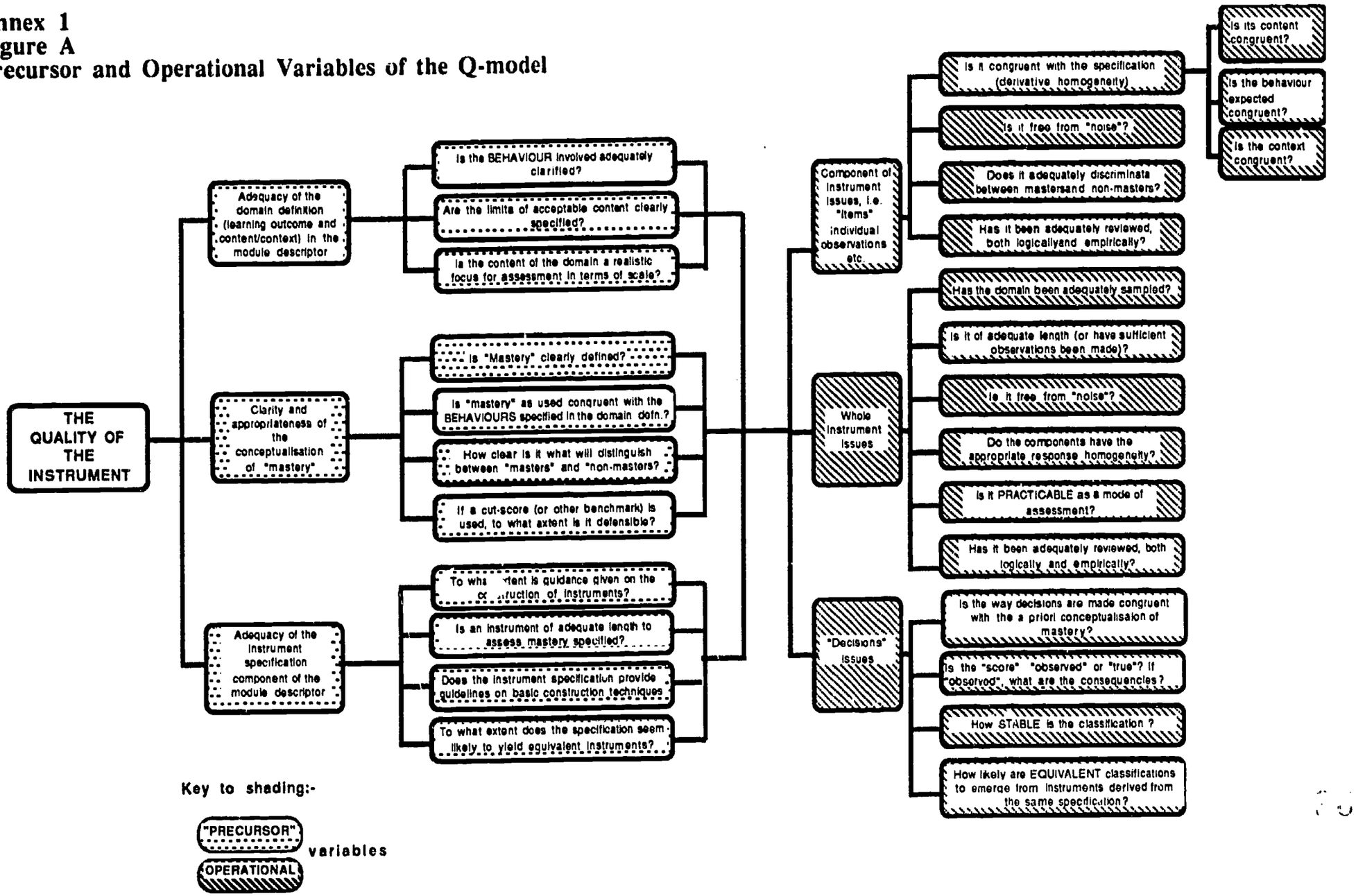
SCOTVEC (no date) *Guidelines on Assessment for Subject Assessors.*

SED (1983) *16-18's in Scotland: an Action Plan.*

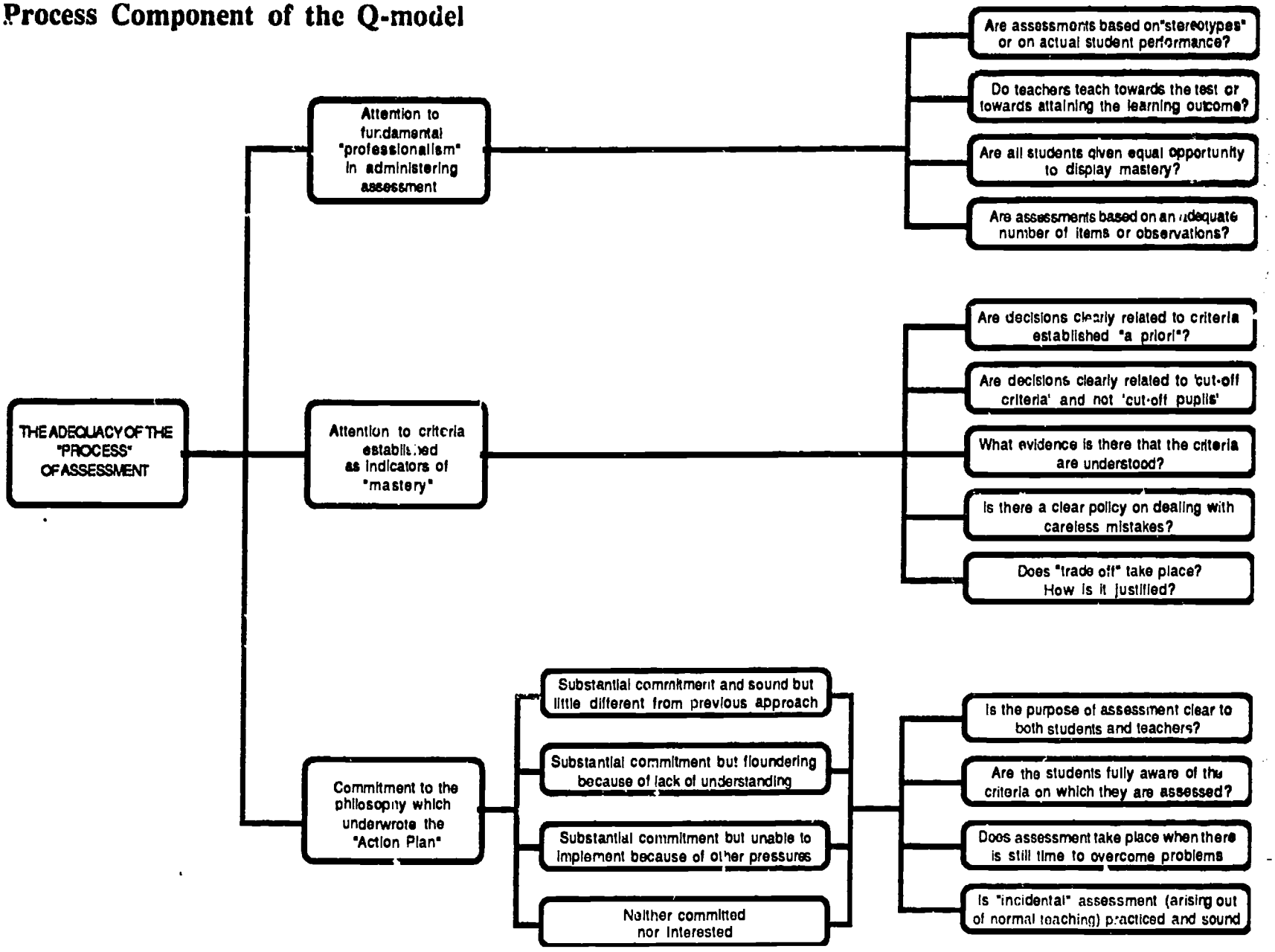
SED (1988) *The National Certificate 1985-86*: Statistical Bulletin 2/F5/1988. SED.

SIEGEL, S (1956) *Nonparametric Statistics for the Behavioural Sciences.* London: McGraw-Hill.

Annex 1
Figure A
Precursor and Operational Variables of the Q-model



Annex 1
Figure B
The Process Component of the Q-model



GLOSSARY

aggregate score: an overall score obtained by totalling individual scores. Scores for elements of a learning outcome can be aggregated to give an overall score for the learning outcome.

alternative instruments: two assessment instruments which are designed to assess the same domain. They should be both *functionally* and *derivatively homogeneous* and should thus result in the same decision regarding the student's mastery/non-mastery status.

assessment specifications: see *instrument specifications*.

comparability: the extent to which two assessment instruments or assessors produce the same, or similar, assessment decisions. Comparability can be sub-divided into *equivalence* and *stability*.

derivative homogeneity: the extent to which an item or an instrument is an adequate reflection of the domain as defined in the descriptor.

diagnostic assessment: assessment which is designed to identify a student's strengths and weaknesses with the purpose of enhancing his or her learning.

domain definition: that area of knowledge, skill or behaviour which is the subject of assessment. In the National Certificate the domain definition is given by the learning outcome, performance criteria and content/context.

domain score: the score that a student would obtain if he or she answered all the possible items which could be generated to assess that domain. Since the number of such possible items is often infinite, this is a purely theoretical construct. See *true score*.

domain score estimate: a score obtained on a sample of items assessing a domain. A 'best guess' at what the *domain score* might be.

empirical review: statistical analysis of students' performances on an assessment instrument for one or more of the following reasons:

* to highlight defective items
* to evaluate the test as a whole
* to examine the degree of *comparability* between two or more assessments.

Empirical review is carried out after an assessment instrument has been used to assess students.

equivalence: the degree of *comparability* between two assessment instruments or two assessors assessing the same students at (approximately) the same time. See *stability*, *reliability*.

formative assessment: assessment used in the course of teaching as an aid to learning. Not necessarily recorded for *summative* purposes.

functional homogeneity: the degree to which two assesments can be said to be 'doing the same job', ie assessing the same things and enabling assessors to reach the same decisions regarding students' performances.

instrument specifications:the details of what type of assessment instrument to use, the format in which it should be presented and other technical considerations which must be addressed when writing assessment items.

logical review: inspection of an assessment instrument or items by subject or assessment experts to ensure that it matches the *domain definition* and *instrument specifications*. It is usually carried out before the instrument is used.

'loosely' defined modules: modules in which, for one reason or another, the *domain definition* and *instrument specifications* are not, or cannot be, clearly defined in advance. See *'tightly' defined modules*.

mode (language): in language studies it is common to talk of the four 'modes' of reading, writing, talking and listening. In the Communication modules these are equivalent to the four learning outcomes.

observed score: a student's actual score on an assessment. See *true score*.

operational variables: in the Q-model devised in this research the operational variables are those factors which come into play when an assessment instrument has been devised and is administered to students. See *precursor variables*, *process variables*.

precursor variables: in the Q-model the precursor variables are all those prerequisite factors which must be taken into account before an assessment instrument is constructed. In practice this means the information provided by the module descriptor. See *operational variables*, *process variables*.

process variables: in the Q-model the process variables are those factors which relate to the way in which an assessor conducts an assessment or uses an assessment instrument. See *operational variables*, *precursor variables*.

Q-model: the model devised by this project to identify the major factors which influence the quality of assessment in the National Certificate. These are grouped into *precursor*, *operational* and *process variables*.

reliability: the technical term for the degree of consistency or *comparability* of assessments. Reliability comes in many forms and is a difficult term to use without further qualifying its meaning in a particular context. The term *comparability* is preferred in this report.

stability: stability refers to comparability over time. If a group of students is assessed twice using the same assessment instrument, but with some time between the two assessments, then stability is the extent to which the same results are obtained on the two occasions.

summative assessment: an assessment conducted to measure the final level of achievement of students, usually for certification purposes.

'tightly' defined modules: modules in which the *domain definition* and *instrument specifications* are clearly and unambiguously defined in advance, leaving little scope for subjectivity or individual judgement.

true score: strictly speaking, the score which a student would have obtained for a domain, were he or she to have answered all the possible questions for that domain. This would make it the same as the *domain score*. More usually, it is the *observed score* statistically adjusted to take account of random error, length of test etc.

validity: the extent to which an assessment instrument actually does measure what it purports to measure. There are various forms of validity, depending on whether one is interested in examining if an assessment instrument is focusing on the correct skills, knowledge etc; if it is adequately covering all the relevant aspects of the domain or if it adequately fulfils the purpose for which it is intended.

SCRE
Publication 107
ISBN
0 947833 34 X

SCRE Practitioner Papers have a practical slant. They present research findings and issues clearly and succinctly to help teachers and others take account of educational research in improving the practice of education. Some titles are of most interest to certain groups — headteachers, staff in further education or those concerned with staff development, for example. Others will attract a wider readership but all are written for an identified audience. The series includes reports of research, edited collections around a theme, reviews of research and annotated bibliographies.

The soundness of assessments teachers make of students' performance on National Certificate modules is vital to students and teachers alike. The NC assessment system is radical in its use of criterion-referencing, in the responsibilities it gives to teachers and in the demands it makes for quality control to ensure fairness and comparable gradings. What can be learnt after five years experience of working the system? The second report of the investigation commissioned by the SED into National Certificate assessment focuses on quality -how sound are the assessments made by teaching staff and what influences quality? Through case-studies chosen to include 'practical' and 'academic' subjects the authors attempt to answer these questions and discuss the implications for those assessing and teaching modules.

SCOTTISH COUNCIL FOR RESEARCH IN EDUCATION